energies

MDPI

Article

Grid-Related Fine Action Segmentation Based on an STCNN-MCM Joint Algorithm during Smart Grid Training

Yong Liu [1], Weiwen Zhan [1], Yuan Li [2], Xingrui Li [1], Jingkai Guo [1] and Xiaoling Chen [3,*]

[1] School of Mechanical Engineering and Electronic Information, China University of Geosciences, Wuhan 430074, China
[2] School of Physical Education, China University of Geosciences, Wuhan 430074, China
[3] School of Art and Media, China University of Geosciences, Wuhan 430074, China
* Correspondence: babyvslee1009@cug.edu.cn

Abstract: Smart grid-training systems enable trainers to achieve the high safety standards required for power operation. Effective methods for the rational segmentation of continuous fine actions can improve smart grid-training systems, which is of great significance to sustainable power-grid operation and the personal safety of operators. In this paper, a joint algorithm of a spatio-temporal convolutional neural network and multidimensional cloud model (STCNN-MCM) is proposed to complete the segmentation of fine actions during power operation. Firstly, the spatio-temporal convolutional neural network (STCNN) is used to extract action features from the multi-sensor dataset of hand actions during power operation and to predict the next moment's action to form a multi-outcome dataset; then, a multidimensional cloud model (MCM) is designed based on the motion features of the real power operation; finally, the corresponding probabilities are obtained from the distribution of the predicted data in the cloud model through the multi-outcome dataset for action-rsegmentation point determination. The results show that STCNN-MCM can choose the segmentation points of fine actions in power operation in a relatively efficient way, improve the accuracy of action division, and can be used to improve smart grid-training systems for the segmentation of continuous fine actions in power operation.

Keywords: power-grid training; cloud model; spatio-temporal convolutional neural network; action segmentation

check for updates

Citation: Liu, Y.; Zhan, W.; Li, Y.; Li, X.; Guo, J.; Chen, X. Grid-Related Fine Action Segmentation Based on an STCNN-MCM Joint Algorithm during Smart Grid Training. *Energies* **2023**, *16*, 1455. https://doi.org/10.3390/en16031455

Academic Editor: Ahmed Abu-Siada

Received: 26 December 2022
Revised: 21 January 2023
Accepted: 26 January 2023
Published: 1 February 2023

1. Introduction

In the process of power operations, safety accidents caused by operational errors often occur, threatening the lives of front-line power workers. Increasingly, training which is costly or dangerous in reality is being conducted in a virtual environment. For example, VR training in medical-skills training can improve the technical skills of surgeons in orthopedic surgery, while saving costs [1], and VR training for skills training in hazardous scenarios can ensure the safety of trainers while learning basic fire-safety skills [2]. The smart grid-training system mentioned in this paper evaluates the standardization of trainers' operating actions by conducting simulation experiments on the fine actions of power operation, which improves the precision of power workers' operations and reduces the potential risks of man-made safety accidents.

A smart grid usually refers to the introduction of artificial-intelligence technology in the traditional power grid, in order to achieve intelligent deployment of power resources; intelligent monitoring of power-system security and other functions; and to improve the reliability, efficiency, safety and environmental and energy sustainability of the power-grid system operation [3]. The smart grid-training system is based on AI technology and VR technology to achieve intelligent safety training, which can intelligently evaluate the ability of operators to ensure personal safety and maintain the normal operation of the power-grid system. For the intelligent evaluation of the standardization of actions during power operations, intelligent segmentation of continuous fine actions is needed to analyze each sub-action and then, finally, calculate a comprehensive score.

A prerequisite for the segmentation of fine actions during power operation is that the data required for the experiments is obtained precisely during hand movements. VR technology allows the real-time data monitoring of hand movements, which is an excellent match for the need to obtain highly precise data. For example, calculating the degree of a hand tremor in a virtual environment can be used to determine treatment options for Parkinson's, in reality [4]. This paper is based on artificial-intelligence algorithm and VR technology to improve the smart grid-training system. Detailed data information from hands can be easily and precisely captured through data gloves, as well as the digital processing of target elements such as movements and objects in the virtual training systrm, which is more conducive to extracting the needed features for analysis. A neural network model is constructed to complete adaptive learning and improve the accuracy of action prediction in order to achieve the segmentation of continuous fine actions during power operations by combining with cloud models.

When analyzing the standardization of power operations, it is often necessary to break it into several simple actions for analysis, and the several actions obtained by splitting the continuous action are called sub-actions. The characteristics of each sub-action differ substantially from each other, so it is necessary to perform action segmentation when the difference in characteristics changes during a continuous action, and then evaluate each sub-action. In a smart grid-training system, segmentation of the continuous fine actions that exist during power operation is extremely important. How to decompose the standard actions of power operation in smart grid-training systems and how to evaluate the consistency of the decomposed action behaviors can greatly contribute to subsequent research on action similarity.

The contributions of this paper are as follows: First, this paper proposes applying VR technology in grid-related operation training projects to ensure the personal safety of the trainers, and proposes the use of data gloves to capture data and dataize action features. Second, this paper applies an attention mechanism to optimize spatio-temporal convolutional neural networks to provide a neural network model which efficiently extracts action features for action prediction of power operations. Third, multi-outcome datasets are constructed based on the action prediction results of the spatio-temporal convolutional neural networks, and STCNN-MCM is proposed to find reasonable action-segmentation points in the process of power operation and improve the accuracy of action segmentation.

This paper is organized as follows: Section 1 introduces the grid-based training project using VR technology, and the issue of how to improve the accuracy of action segmentation in a smart grid-training system, and presents the solution of this paper. Section 2 introduces the worldwide research achievements and solutions to related topics, and identifies the problems that have not been solved, presents data gloves and proposes solutions. Section 3 introduces the STCNN used in this paper, and introduces the attention mechanism, the method of constructing the normal cloud model and the action-segmentation logic based on STCNN-MCM. In Section 4, the method of acquiring hand motion data is introduced, and analysis of the available results of the experiments is performed. In Section 5, the novelty of this paper and the limitations and deficiencies of STCNN-MCM are discussed, in comparison with existing work. In this paper, the conclusion that STCNN-MCM can be applied to the segmentation of continuous fine actions in power operation is justified, and it is proven that STCNN-MCM jointly constructed with a spatio-temporal convolutional neural network, attention mechanism and multidimensional cloud model can improve the accuracy rate of action segmentation in a smart grid-training system.

2. Related Works

Action segmentation and action recognition are usually achieved by extracting and learning features, and feature-extraction technology is constantly being updated to become more accurate and efficient [5]. In [6], Wu and Shao proposed a hierarchical dynamic framework which can extract high-level skeletal joint features from 3D skeletal data and used the learned representations to estimate firing probabilities, inferring action sequences by replacing Gaussian mixture models with deep neural networks, while segmenting and identifying actions. In [7], Zhu et al. proposed an online continuous human action-recognition algorithm based on a Kinect sensor, which has high efficiency for the recognition of continuous human behavior. In [8], a temporal convolutional network (TCN) was proposed by Lea et al., which uses a hierarchy of temporal convolutions for fine-grained action segmentation or detection, and is able to capture action composition, segment duration, and long-range dependencies, outperformed recurrent neural networks (RNN) [9] and long short-term memory (LSTM) [10], which are recognized as suitable for processing timing problems, on a variety of tasks. On this basis, multi-stage temporal convolutional networks were proposed by Farha and Gall in [11]; each stage uses a multi-layer null convolution with increasing sensory field layer by layer, and the range of the field of view obtained increases as the number of layers deepens, to better capture information over a prolonged period of time. This multi-frame model is effective in capturing long-range dependencies and identifying action segments. Introducing external information can increase the accuracy of action segmentation, a boundary-aware cascade network (BCN) incorporating action boundary information was proposed by Wang et al. in [12], which uses a multi-stage network structure with dynamic modeling capability and adaptively employs different sub-networks to process samples according to their difficulty level, with shallow sub-networks for simple samples and deep sub-networks for difficult samples, as a way to improve the classification accuracy of difficult samples. In [13], Gao et al. proposed a human-activity classification and recognition model based on smartphone inertial sensor data, which segmented human-activity data and recognized it based on different classifiers. In [14], dos Santos et al. proposed semi-supervised and iterative reinforcement learning (RL-SSI) with the purpose of improving the recognition accuracy of human actions in videos.

CNNs have been used in the field of action recognition and action segmentation in stages [15]. 3DCNN, as an improved CNN algorithm, can extract features from spatial and temporal dimensions by performing 3D convolution, which captures the motion information encoded in multiple adjacent frames and provides a significant improvement in the accuracy of human-action recognition [16]. In [17], Molchanov et al. proposed an algorithm to identify drivers' gestures from depth and intensity data using a 3D convolutional neural network, combining information from multiple spatial scales for prediction, and employing spatial-temporal data augmentation for more efficient training and reducing potential overfitting. In [18], Lea et al. proposed an action-segmentation model to convolve temporal and spatial features separately, which combines low-level spatio-temporal features and high-level segmentation classifiers to complete action segmentation based on captured features through semi-Markov models. Given the ability of 3DCNNs to extract spatio-temporal features within video frames, various resource-efficient 2DCNNs were converted to 3DCNNs and the performance was evaluated. It was found that the performance after conversion to 3DCNNs can be considerably accurate and memory usage in terms of classification accuracy for different levels of complexity is reduced [19]. In [20], Martin et al. proposed an "end-to-end" scheme using deep neural networks to extract features for action-recognition tasks, and 3DCNN can effectively capture spatio-temporal features for action classification from videos. In [21], Chen et al. presented a novel CNN-based motion estimator developed for complex fluid flows using multiscale cost volumes, with state-of-the-art evaluation results on a public fluid dataset.

Many studies have optimized deep-learning models by adding attention modules, and attention mechanisms have become one of the most important concepts in the field of deep learning [22]. In [23], an attention module based on the spatial domain and channel domain is proposed by Kim et al., and the attention module is based on a triplet loss function to distinguish the action part from the non-action part. In [24], Zhu et al. proposed a spatio-temporal graph modeling method to build a spatio-temporal graph attention network

(STGATP) for traffic prediction, which captures both spatial and temporal dependencies in the road network and achieves excellent prediction performance for traffic problems. In [25], He et al. proposed a vector graph convolution deep-learning model optimized for virtual hand data to modify the spatial graph convolutional layer and introduce a graph attention module to improve the accuracy of action recognition in virtual environments. In [26], a sea surface temperature (SST) prediction method based on 3DCNN and LSTM networks was proposed by Qiao et al., adding an attention mechanism to weight the output of each step of the LSTM model to improve the prediction accuracy. In [27], Wei et al. proposed a TPA-LSTM hybrid model for predicting future ridership at subway stations which is based on the TPA mechanism and LSTM networks.

Cloud models belong to the category of AI with uncertainty and are mainly used for interconversion between qualitative and quantitative measurements [28], and have been used in several fields. In [29], a generic multidimensional cloud model was proposed by Deng et al., which can reasonably characterize spatial entities, reveal the spatial distribution of potential information, and achieve more accurate spatial delineation in complex situations based on the ideas of non-homogeneity and non-symmetry. The multidimensional similarity cloud model (MSCM) was combined with a stochastic weighting method to reduce the impact of random errors in eutrophication monitoring data and ambiguity in the definition of lake eutrophication on the consistency and reliability of lake-eutrophication evaluation [30]. In [31], Wang et al. proposed a novel multidimensional connected cloud model to describe the uncertainty and distribution characteristics of indicators, as well as the fuzziness of classification boundaries, so as to predict rock bursts with numerous random and fuzzy indicators. In [32], Zhang and Liu proposed a HHCE-MOPSO hybrid model based on a normal cloud model and entropy, and verified the feasibility of the hybrid particle-swarm algorithm.

In the above, many existing methods are able to extract action features from video or continuous frame images. However, there are few procedures for processing data obtained based on data gloves to segment continuous fine actions in power operation. In grid-related training based on VR technology, human–machine interaction and data acquisition are performed by the data glove, which is shown in Figure 1. Although cloud models are widely used in several fields, there is no research on applying multidimensional cloud models to action recognition and action segmentation in virtual environments.

Figure 1. VR data glove.

In response to the problems of the above studies, data is captured with data gloves in a virtual-reality environment in this paper. Firstly, the data is preprocessed, a hand spatial-temporal data matrix is constructed, and STCNN-MCM completes the segmentation of continuous fine actions in power operation.

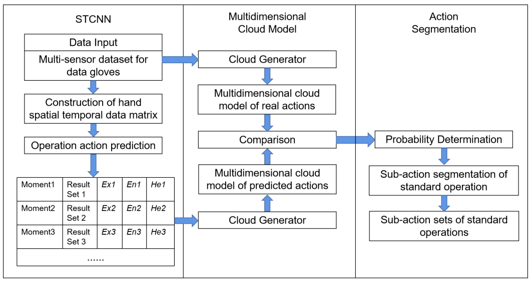
3. System Model

Human action segmentation is an active research area in computer vision and has a wide range of applications in major fields. However, traditional action segmentation often relies on hand-extracted features, and these methods tend to ignore many high-level features that are embedded in the hand structure. Dynamic hand movements convey significant information; the segmentation of hand movements in grid-related operation training requires a higher degree of accuracy.

Therefore, this paper proposes STCNN-MCM for solving some of the problems of traditional action segmentation and improving the accuracy of action segmentation. Firstly, the hand action data is obtained during the operation through the data gloves, and the hand spatial-temporal data matrix is constructed based on the hand-structure sequence; the convolutional layer is then modified to add the attention mechanism to complete the training of STCNN, and the multi-result dataset constructed based on the prediction results of STCNN. Finally, the corresponding probability is obtained through the distribution of the predicted data in the cloud model in the multi-result dataset for the action-segmentation points, which are determined to complete the segmentation of the testers' actions.

3.1. *Joint Algorithm Design*

In this paper, a database was constructed by applying the hand information data which was received from the data gloves, and the segmentation of hand movements based on the dataset was completed for continuous fine actions to assist the action evaluation system later. Thus, the grid operators can correct irregular operations during their own training. Unreal Engine is an open and advanced real-time 3D authoring tool which provides realistic visuals and immersive experiences, and the entire smart-grid training system was built on UE4. By collaborating with industry-related professional practitioners, UE4 was used and blueprints were programmed for glove data collection of the professionals' hand operation data. For readability considerations, the hand information data were recorded in an XML format with the hand nodes, and preprocessed to process one frame of data into a two-dimensional matrix. The training data sets of different lengths were set for the STCNN based on the action data from the previous moment of the operation, and the prediction of the hand action at the latter moment could be completed based on the obtained training model to build a prediction set for each moment. The prediction result set of each moment was constructed based on the obtained training model. All the prediction datasets together constitute a multi-outcome dataset, and the multidimensional cloud model constructed based on the prediction results was compared with the multidimensional cloud model constructed based on the real actions, to obtain the probability that the next moment of movement would conform to the degree of human motion change. The motion features in the prediction result set are matched in the cloud space, and when the data features of the predicted action had a probability lower than a certain value in the space, the corresponding moment was used as the termination decision of the sub-action and at the same time as the start of the next sub-action. This start and end determination was used to partition the overall action into actions and obtain a standard set of sub-actions. The refined flow chart of the joint algorithm for the segmentation of hand actions for continuous fine actions is shown in Figure 2.



STCNN
Data Input
Multi-sensor dataset for data gloves
Construction of hand spatial temporal data matrix
Operation action prediction
Moment1
Result Set 1
Ex1
En1
He1
Moment2
Result Set 2
Ex2
En2
He2
Moment3
Result Set 3
Ex3
En3
He3
......
Multidimensional Cloud Model
Cloud Generator
Multidimensional cloud model of real actions
Comparison
Multidimensional cloud model of predicted actions
Cloud Generator
Action Segmentation
Probability Determination
Sub-action segmentation of standard operation
Sub-action sets of standard operations

Figure 2. Joint-algorithm-refined flow chart.

3.2. *Spatio-Temporal Convolutional Neural Network*

For the action segmentation of the grid class, the input data features to be identified are continuous, with multiple operational video frames of data containing temporal features;

therefore, a spatio-temporal convolutional neural network (STCNN) was used to extract data features in temporal and spatial terms.

STCNN is convolved with a three-dimensional convolution kernel in both the temporal dimension and the spatial dimensions of width and height, so that the feature maps acquired by convolution contain spatial features of multiple consecutive frames, which means that the spatial features with a temporal dimension are learned. Three-dimensional convolution is achieved by convolving a three-dimensional kernel with a cube formed by stacking multiple consecutive frames together.

In 2DCNN, the convolutional layer performs 2D convolution to extract features from local neighborhoods on the feature maps of the previous layer. A 2D convolution is performed on the image, and the value v of the (x, y) of the j-th feature map of the i-th layer can be derived from the following formula.

$$[conv(a,w)]_{a_{ij}}xy = tanh(b_{ij} + \sum_m \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} w_{ijm}^{pq} a_{(i-1)m}^{(x+p)(y+q)}), \tag{1}$$

where $tanh$ denotes the hyperbolic tangent function, which represents the bias of this feature map; m is the index between the $(i-1)$-th layer feature map and the current feature map; w_{ijm}^{pq} is the value of the (p, q) position in the kernel connected to the m-th layer feature map; P_i is the height of the kernel; and Q_i is the width of the kernel.

STCNN can calculate features from spatial and temporal dimensions, and the value of (x, y, z) in the j-th feature map of the i-th layer can be derived from the following formula.

$$[stconv(a,w)]_{a_{ij}}xyz = tanh(b_{ij} + \sum_m \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} \sum_{r=0}^{R_i-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)}), \tag{2}$$

where R_i is the size of the 3D kernel in the time dimension and w_{ijm}^{pqr} is the value of the (p, q, r) position in the kernel connected to the previous layer of the feature map. The STCNN structure used in this paper is shown in Figure 3.

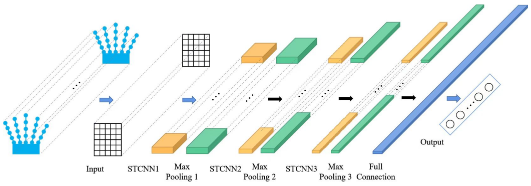


Figure 3. STCNN model structure.

The process of training the STCNN involves optimizing the parameters W of the network to minimize the cost function of the data set D. Negative log likelihood is used to construct the cost function.

$$T(W, D) = -\frac{1}{|D|} \sum_{i=0}^{|D|} \log(P(R^i|v^i, W)). \tag{3}$$

After completing the optimization of the parameter W, the trained STCNN model is obtained. Using the n-frame hand spatial-temporal data matrix as the input of the model, the action feature of the $n+1$ frame can be predicted.

3.3. Construction of Hand Spatial Temporal Data Matrix

After determining the use of STCNN, the hand spatial-temporal data matrix is required to be constructed for this model as input. The hand spatial data matrix needs to be constructed based on the hand-structure map of the data glove in order to be able to express the motion characteristics of the whole hand movement. The construction of the whole hand spatial-temporal data matrix can be performed in two steps. First, all the hand-joint point data is visualizes in each frame sequence and the hand spatial data matrix of the current frame constructed; second, for different joint points of the hand, the same joint points are connected in consecutive frames to construct the hand spatial data matrix of consecutive frames. The construction process of the hand spatial-temporal data matrix is shown in Figure 4.

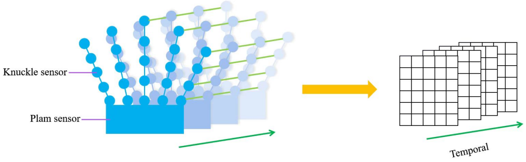
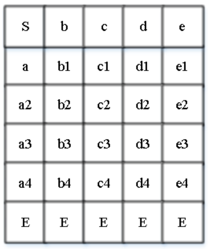


Figure 4. Diagram of the process of constructing the hand spatial-temporal data matrix.

Firstly, the spatial data matrix of the hand for each frame of action is constructed. In the space, the hand wrist is connected to each finger, and the joints at the wrist are kept relative to each joint of the finger. In this paper, the model hand data are visualized by setting the representative letters from thumb to little finger as a, b, c, d, and e. The corresponding joints are represented in the form of a2, b1, c3, d4, etc. As the thumb lacks a node shared with the other fingers, there is no a1 in the matrix. The first of the two nodes of the joint at the wrist is set to the thumb as the starting item of the data, and is denoted by S (start), and the last node ending item is denoted with E (end). The joints of the finger roots are adjacent to the wrist, and the joints of the remaining corresponding fingers are adjacent to the joints of the adjacent fingers in the space position. Among them, the joints of the thumb and pinky fingers are located at the edge of the space, and the joints of index, middle, ring and oinky fingers are located at the top of the palm in space. A frame of data is processed into a two-dimensional matrix with a size of 5*6 and thirty nodes, as shown in Figure 5. The data thus processed maintains the relative position of each finger joint in space, from left to right is the distribution of joints according to the order of the virtual hand model, and from top to bottom is the adjacent distribution according to the joint connection of each finger of the virtual hand. The bottom layer is set with multiple end joints, which is to ensure that the joints near the wrist of each finger can be directly adjacent to the wrist joints, so it can be effectively ensured that each finger is directly adjacent to the wrist. In addition, for each finger, it is only necessary to consider the information of each joint in the space.



S	b	c	d	e
a	b1	c1	d1	e1
a2	b2	c2	d2	e2
a3	b3	c3	d3	e3
a4	b4	c4	d4	e4
E	E	E	E	E

Figure 5. One-frame hand spatial data matrix.

The data glove, during motion, records the motion data of each joint node, and the data includes the three values of the coordinate offset, rotation angle, and scaling factor of that joint in the virtual space. The focus of this study is to consider the information of the virtual hand's position change in the virtual space. After placing the information of the spatial coordinates in the matrix in Figure 5, a 5*6 matrix is obtained, which is the form of the original data after pre-processing. In a segment of hand action, 10 frames are extracted by taking 1 frame at a certain frame interval, and the same nodes of the hand spatial data matrix of these 10 frames are connected in temporal order to form the hand spatial-temporal data matrix.

The input training data set is determined according to the operation segments of the power operation; the previous part of the data which does not affect the real operation is used as the first set of input training data in this paper. During training, the operation actions are divided into multiple groups according to time frames, and the first set of input data predicts the next frame or the last frame of this set of data, i.e., the next frame of all data input for training. The training data is set up by first taking the whole set of data as input, and then decreasing the first frame, the second frame, and up to the Nth frame, in turn. By setting the training data in this way, the predicted frames can be fully considered as input, and reducing the data far from the predicted frames can strengthen the influence of the closer action data on the model parameters, so as to better optimize the model. The convolution process of the hand spatial-temporal data matrix is shown in Figure 6.

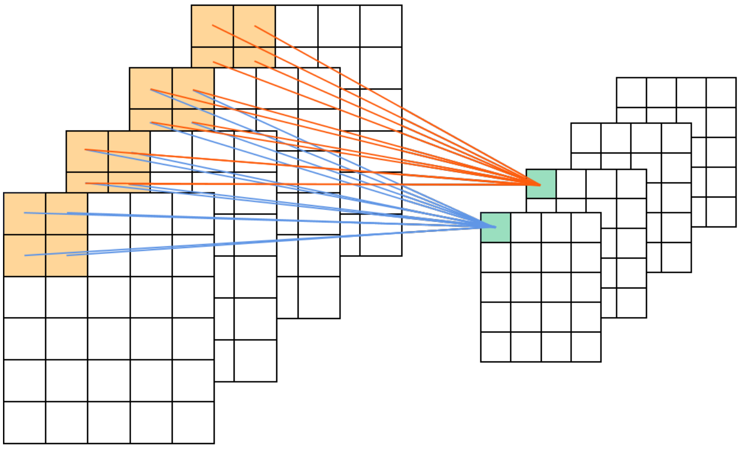


Figure 6. Hand spatial-temporal data matrix convolution process diagram.

3.4. Attention Mechanism

During movement, different trunks are important in different ways. For example, the movement of the legs may be more important than the neck, and jogging, rising and sitting can be distinguished by the legs, whereas the movement of the neck could contain little valid information; therefore, various weights for different trunks can be assigned. In this paper, the attention module will be added to weight some key hand skeleton nodes to make the hand spatial information predicted by STCNN more accurate.

The motion of the entire wrist and arm is relative to the spatial position of the overall human model, and more attention should be paid to the wrist during the motion. The convolutional layer is modified to introduce a spatial attention module to increase the weight of spatial information of key nodes such as wrist joints. The processed data can ensure the overall motion information and the feature information of each joint point comprehensively, and the spatial attention module is shown in Figure 7.

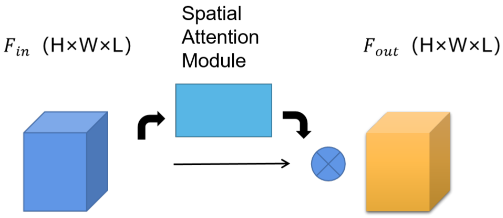


Figure 7. Spatial attention module.

The original STCNN implementing feature extraction in the spatial dimension is given by:

$$f_{out} = \sum_i^k w_i(f_{in}(A_i \odot M_i)), \quad (4)$$

where f_{in} denotes the input to the convolutional layer of a normal STCNN, and f_{out} denotes the output of the input data after convolution; K denotes the size of the convolution kernel; A is the matrix normalized by the hand space information matrix; M is a learnable weight matrix; and the symbol $\odot$ denotes the dot product.

With the addition of the attention module, the convolution formula in the spatial dimension can be expressed as:

$$F_{out} = \sum_i^k w_i(F_{in}(A_i' + B_i)), \tag{5}$$

where F_{in} denotes the input of the convolutional layer with the addition of the spatial attention module, and F_{out} denotes the output of the input data after convolution; A' is the spatial information matrix obtained by the joint action of A and the weight matrix M; and B denotes the attention matrix, which helps the model to better target each sample for feature extraction.

The attention module calculates the similarity between the two based on the current target features and all source features at all times by means of a scoring function, which is as follows:

$$S(F_t, F_k) = v_a^T tanh(W_a[F_t; F_k]). \tag{6}$$

A simple neural network is constructed to obtain a learnable two-dimensional parameter matrix and a one-dimensional parameter vector to calculate the score. In this paper, two full connection layers were used to implement this network in the scoring function, and the alignment weights can be calculated by the softmax function after the scores are obtained, calculated as follows:

$$a_t(k) = \frac{\exp(S(F_t, F_k))}{\sum_{k'} \exp(S(F_t, F_{k'}))}. \tag{7}$$

The alignment weight a_t is the similarity between the target feature F_t and all source features F_k, which is the element of matrix B in Formula (5). It can be seen that the matrix B is also obtained entirely by learning different action samples, and it can effectively learn the weights of any hand joints in different actions, which increases the flexibility and generality of the model and makes the model effective in action prediction even in the face of diverse data.

By introducing an attention mechanism into the convolutional layer, this method can effectively extract effective joint information from hand movements and improve the accuracy of action prediction.

3.5. Construction of Normal Cloud Model

The numerical characteristics of the cloud include expectation Ex, entropy En and hyperentropy He, which are the parameters used for the quantitative analysis of hand sub-actions in continuous fine actions. Expectation, Ex, is the expected distribution of cloud droplets in the domain space from which the qualitative concept of a certain stage of a hand sub-action in a continuous fine actions can be quantitatively defined. Entropy, En, is a measure of the uncertainty associated with the qualitative concept of a particular stage of a sub-action. Entropy represents the degree of the dispersion of cloud droplets and the range of cloud droplets in the domain space which can be accepted as belonging to each hand sub-action. Thus, entropy can be used to effectively represent the qualitative and inherently ambiguous concepts of hand sub-actions. The super entropy He is a measure of the uncertainty associated with entropy, which reflects the degree of cohesion between cloud drops. A schematic representation of the numerical characteristics of a normal cloud model is shown in Figure 8.

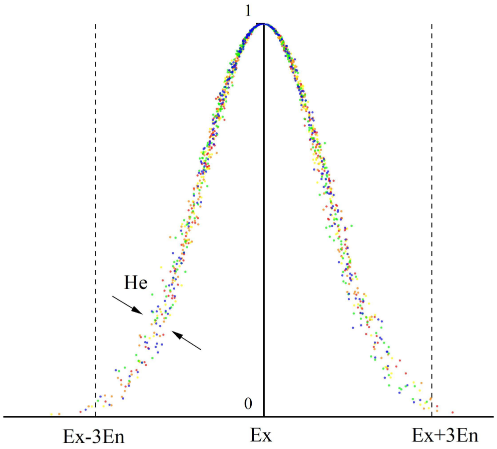


Figure 8. Cloud-model digital features schematic.

By taking the action characteristics of hand sub-actions at a certain stage as the input of the forward cloud generator, the affiliation degree scores of hand sub-actions at that stage can be calculated, and the analysis of hand sub-actions at that stage from qualitative to quantitative can be completed. On the contrary, using the random hand-action sample data as the input of the reverse cloud generator, the cloud digital features of that hand action can be calculated, and the analysis of the hand sub-action at that stage, from quantitative to qualitative, can be completed.

In this paper, the normal cloud model is constructed by the forward cloud generator to analyze the characteristics of hand sub-actions from qualitative to quantitative, so as to complete the hand action segmentation for continuous fine actions. The construction process of the forward cloud generator is as follows.

(1) Determination of *Ex*

For a variable which has a certain range of variation, shaped as $V_{Qa}[B_{min}, B_{max}]$, Ex can generally be calculated by the following formula.

$$Ex = (B_{min} + B_{max})/2, \tag{8}$$

where B_{min} and B_{max} denote the minimum and maximum boundaries of the variable V_{Qa}. For a variable with unilateral boundaries, shaped as $V_{Qa}[B_{min}, +\infty]$ or $V_{Qa}[-\infty, B_{max}]$, default boundary parameters or expected values can be determined first based on the upper and lower boundaries of the variable, and then the numerical characteristics of the cloud Ex calculated according to Formula (8).

(2) Determination of *En*

Since the established multidimensional cloud model has to consider all the changes in each variable comprehensively, the value of change in each variable in this paper, to determine the numerical characteristics of the cloud En according to the maximum range of its change, and the En of this evaluation factor were kept constant, as determined by the following formula.

$$En = (Ex)/3, \tag{9}$$

where Ex is the different expected values corresponding to a certain variable. The formula here is set according to the $3Ex$ rule.

(3) Determination of *He*

He, the numerical characteristic of the cloud, can be chosen as a suitable constant k according to the maximum range of each evaluation factor, in general. If the distance between cloud drops is too large as well as scattered, the qualitative concept is not well-expressed. *He* is calculated as follows.

$$He = kEn. \tag{10}$$

(4) Forward cloud generator

The three numerical characteristics of the cloud are input and the affiliation degree scores calculated to determine the distribution of cloud drops. The schematic diagram of the forward cloud generator is shown in Figure 9.

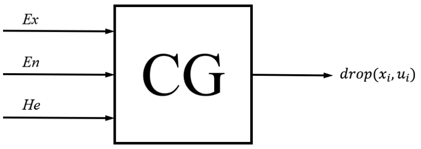


Figure 9. Schematic diagram of the forward cloud generator.

The affiliation degree score is calculated by the following formula.

$$\mu_i = \exp[-\frac{1}{d}\sum_{j=1}^{d}\frac{(x_{ji} - Ex_j)^2}{y_{ji}^2}], \tag{11}$$

where d denotes the maximum dimensionality of the generated cloud model, x_{ji} is the i-th j-dimensional normal random number generated with $(Ex_1, Ex_2, \ldots, Ex_j)$ as expectation and $(En_1, En_2, \ldots, En_j)$ as variance, and y_{ji} is the i-th j-dimensional normal random number generated with $(En_1, En_2, \ldots, En_j)$ as expectation and $(He_1, He_2, \ldots, He_j)$ as variance. Finally, let (x_i, μ_i) be the cloud drops.

3.6. STCNN-MCM Joint Algorithm

Fuzziness and stochasticity can be integrated by the cloud model. By upgrading the traditional fuzzy and probabilistic theories, the qualitative concept of operational action changes conforming to the operational motion characteristics is converted with the quantitative numerical results obtained from STCNN prediction, and the actions of specified tasks in power operations are segmented, thus making the data glove a better data application.

The result set after performing STCNN prediction is diversity. If the change in the motion position of the virtual hand in accordance with the change in the operational action is considered as a qualitative concept, each outcome in the predicted result set can be considered as a stochastic realization of this qualitative concept in the quantitative domain. The spatial location of each result expresses the degree to which the predicted result supports the qualitative concept of conformity to operational changes in the operation. Therefore, this paper adopts a cloud model to deal with the relationship between quantitative-prediction result distribution and qualitative human-hand motion characteristics, and designs STCNN-MCM to segment grid-related continuous fine actions.

STCNN-MCM consists of the following main steps. 1. $T_k, \ldots, T_{k+m}$, m frames of data as the input of STCNN to predict the action features of T_{k+m+1} frames. 2. n predictions are made, and the prediction results $T_{k+m+1}, \ldots, T_{k+m+1+n}$ for a total of n frames together form the set of prediction results as the action features for the next moment. 3. Cloud transformation is used to convert the motion features in the prediction-result set into numerical features of clouds. 4. The affiliation degree scores calculated from all the predicted motion results obtained from each cloud to determine the attribution of each prediction result to the cloud. 5. For each prediction result set, the forward cloud generator is used to generate the same number of cloud drops to construct a multidimensional cloud model. 6. The multidimensional cloud model constructed from the predicted dataset is compared with the multidimensional cloud model constructed from the real action data to obtain the probability of the predicted motion position in the real action. The construction process of the multidimensional cloud model for predicting actions is shown in Figure 10.

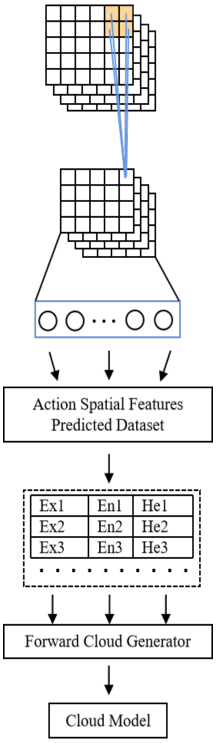


Figure 10. Flowchart of cloud model generation for predicting actions.

4. Application Instances

This paper conducts repeated virtual operation experiments for power operations with more risk points to reveal safety issues in the process of power operations. The electrical inspection is a continuous fine action, which is difficult to match with the standard action directly, and the spatial information of each hand node is extremely important, so it needs to be divided when evaluating such actions, and after the division, the sub-actions can be evaluated more finely, so as to complete the standard evaluation of such actions as a whole. The flow chart of smart grid-training system operation is shown in Figure 11.

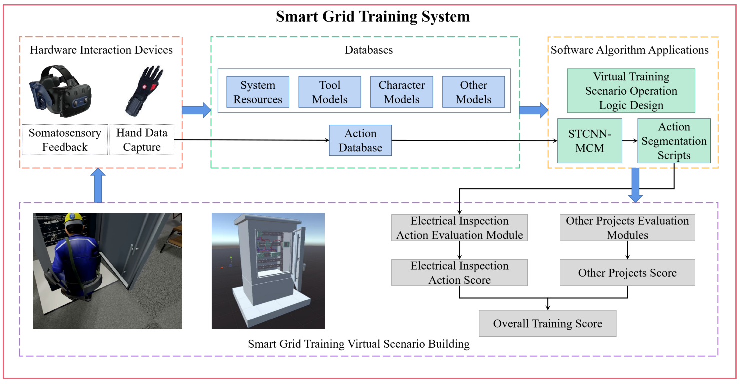
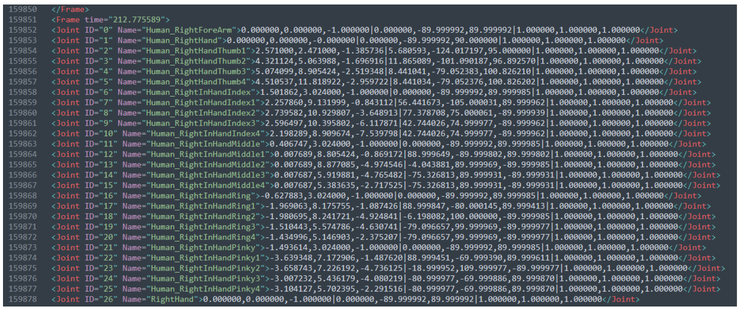

Smart Grid Training System
Hardware Interaction Devices
Somatosensory Feedback
Hand Data Capture
Databases
System Resources
Tool Models
Character Models
Other Models
Action Database
Software Algorithm Applications
Virtual Training Scenario Operation Logic Design
STCNN-MCM
Action Segmentation Scripts
Electrical Inspection Action Evaluation Module
Other Projects Evaluation Modules
Electrical Inspection Action Score
Other Projects Score
Overall Training Score
Smart Grid Training Virtual Scenario Building

Figure 11. Smart-grid-training-system operation flow chart.

4.1. Action-Data Acquisition Method

The standard action data were established by skilled operators of the power grid, and after operating in the training system, a set of these actions was selected that could be used as the judging standard. The whole operation process took about 23 s and a total of 678 frames of data were obtained. Each frame of data included three values of coordinate offset, rotation angle, and scaling factor in virtual space for 27 nodes (24 finger nodes, 1 wrist node, 1 hand center node, 1 node representing the entire hand model), which could be obtained based on the virtual glove. Some of the data for the standard action are shown in Figure 12.



```
</Frame>
<Frame time="212.775589">
<Joint ID="0" Name="Human_RightForeArm">0.000000,0.000000,-1.000000|0.000000,-89.999992,89.999992|1.000000,1.000000,1.000000</Joint>
<Joint ID="1" Name="Human_RightHand">0.000000,0.000000,-0.000000|0.000000,-89.999992,90.000000|1.000000,1.000000,1.000000</Joint>
<Joint ID="2" Name="Human_RightHandThumb1">2.571000,2.471000,-1.385736|5.680593,-124.017197,95.000000|1.000000,1.000000,1.000000</Joint>
<Joint ID="3" Name="Human_RightHandThumb2">4.321124,5.063988,-1.696916|11.865089,-101.090187,96.892570|1.000000,1.000000,1.000000</Joint>
<Joint ID="4" Name="Human_RightHandThumb3">5.074099,8.905424,-2.519348|8.441041,-79.052383,100.826210|1.000000,1.000000,1.000000</Joint>
<Joint ID="5" Name="Human_RightHandThumb4">4.510537,11.818922,-2.959722|8.441034,-79.052376,100.826202|1.000000,1.000000,1.000000</Joint>
<Joint ID="6" Name="Human_RightInHandIndex">1.501862,3.024000,-1.000000|0.000000,-89.999992,89.999985|1.000000,1.000000,1.000000</Joint>
<Joint ID="7" Name="Human_RightInHandIndex1">2.257860,9.131999,-0.843112|56.441673,-105.000031,89.999962|1.000000,1.000000,1.000000</Joint>
<Joint ID="8" Name="Human_RightInHandIndex2">2.739582,10.929807,-3.648913|77.378708,75.000061,-89.999939|1.000000,1.000000,1.000000</Joint>
<Joint ID="9" Name="Human_RightInHandIndex3">2.596497,10.395802,-6.117871|42.744026,74.999977,-89.999962|1.000000,1.000000,1.000000</Joint>
<Joint ID="10" Name="Human_RightInHandIndex4">2.198289,8.909674,-7.539798|42.744026,74.999977,-89.999962|1.000000,1.000000,1.000000</Joint>
<Joint ID="11" Name="Human_RightInHandMiddle">0.406747,3.024000,-1.000000|0.000000,-89.999992,89.999985|1.000000,1.000000,1.000000</Joint>
<Joint ID="12" Name="Human_RightInHandMiddle1">0.007689,8.805424,-0.869172|88.999649,-89.999802,89.999802|1.000000,1.000000,1.000000</Joint>
<Joint ID="13" Name="Human_RightInHandMiddle2">0.007689,8.877085,-4.974546|-4.043881,89.999969,-89.999985|1.000000,1.000000,1.000000</Joint>
<Joint ID="14" Name="Human_RightInHandMiddle3">0.007687,5.919881,-4.765482|-75.326813,89.999931,-89.999931|1.000000,1.000000,1.000000</Joint>
<Joint ID="15" Name="Human_RightInHandMiddle4">0.007687,5.383635,-2.717525|-75.326813,89.999931,-89.999931|1.000000,1.000000,1.000000</Joint>
<Joint ID="16" Name="Human_RightInHandRing">-0.627883,3.024000,-1.000000|0.000000,-89.999992,89.999985|1.000000,1.000000,1.000000</Joint>
<Joint ID="17" Name="Human_RightInHandRing1">-1.969063,8.175755,-1.087426|88.999847,-80.000145,89.999413|1.000000,1.000000,1.000000</Joint>
<Joint ID="18" Name="Human_RightInHandRing2">-1.980695,8.241721,-4.924841|-6.198082,100.000000,-89.999985|1.000000,1.000000,1.000000</Joint>
<Joint ID="19" Name="Human_RightInHandRing3">-1.510443,5.574786,-4.630741|-79.096657,99.999969,-89.999977|1.000000,1.000000,1.000000</Joint>
<Joint ID="20" Name="Human_RightInHandRing4">-1.434996,5.146903,-2.375207|-79.096657,99.999969,-89.999977|1.000000,1.000000,1.000000</Joint>
<Joint ID="21" Name="Human_RightInHandPinky">-1.493614,3.024000,-1.000000|0.000000,-89.999992,89.999985|1.000000,1.000000,1.000000</Joint>
<Joint ID="22" Name="Human_RightInHandPinky1">-3.639348,7.172906,-1.487620|88.999451,-69.999390,89.999611|1.000000,1.000000,1.000000</Joint>
<Joint ID="23" Name="Human_RightInHandPinky2">-3.658743,7.226192,-4.736125|-18.999952,109.999977,-89.999977|1.000000,1.000000,1.000000</Joint>
<Joint ID="24" Name="Human_RightInHandPinky3">-3.007232,5.436179,-4.080219|-80.999977,-69.999886,89.999870|1.000000,1.000000,1.000000</Joint>
<Joint ID="25" Name="Human_RightInHandPinky4">-3.104127,5.702395,-2.291516|-80.999977,-69.999886,89.999870|1.000000,1.000000,1.000000</Joint>
<Joint ID="26" Name="RightHand">0.000000,0.000000,-1.000000|0.000000,-89.999992,89.999992|1.000000,1.000000,1.000000</Joint>
```

Figure 12. Standard-action data display.

From the collected data, it can be seen that the last three scaling factors of each joint are 1. These values indicate that the model constructed by the data glove has a 1:1 ratio to the real hand when the operation is carried out, and this part of the data is eliminated during pre-processing. The study in this paper aims to judge the operation process of electrical inspection; the finger joint angle changes very little after picking up the electric test pen until the pen is put down, so the rotation angle data during the operation has a small impact on the overall movement. During the virtual motion, no activities such as finger rotation occurred during the behavior of the motion, and the significance of studying the rotation angle data is not significant. The first group of three digits is to represent the movement of x, y, z of each joint, respectively, and the unit of data during the movement is 0.001 m.

As shown in Figure 13, below, the offset of the human right-hand middle joint on the x-axis changes. In the virtual space of the movement process, joints need to first move in the counterclockwise direction, in the early move to the tool table to pick up the apparatus faster, pick up when the offset becomes smaller, and then slowly move to the distribution box for power testing operations; the x-direction offset becomes smaller and there is a smaller moving distance during electricity testing. Finally, returning to the tool table, the horizontal offset is negative, and the initial direction of movement is opposite, in line with the law of motion. This shows that the experimental data obtained from the overall operation is reliable.

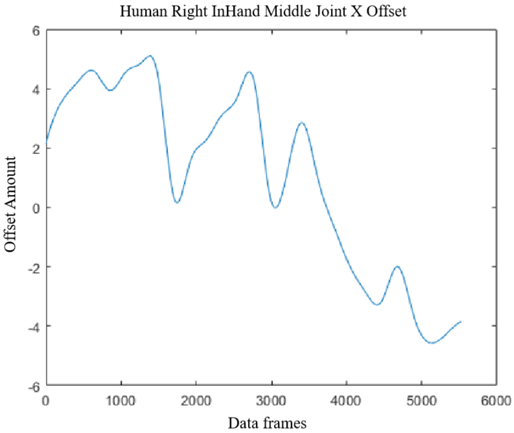


Figure 13. Human right-hand middle joint x offset.

4.2. Action Segmentation of Standard Electricity Inspection

The continuous fine actions studied in this paper constitute the electrical inspection process. The overall process is: picking up the electron microscope and then transporting it to the electrical box waiting to be inspected. The operation in the smart grid-training system is shown in Figure 14.

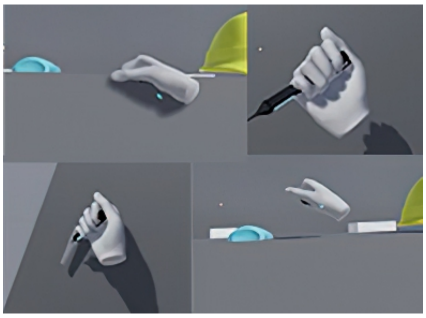


Figure 14. Electrical inspection in smart grid-training system.

After the action of the power operation training combined with the summary analysis of the operating instructions, experts believe that when the human hand carries out a movement to being stationary, maintaining a certain range of motion can be regarded as the operating movement of the human-hand change characteristics. For example, the human hand, maintaining a horizontal movement, may be offset in all directions; when the offset angle is less than 30° it can still be regarded as a movement in line with operational changes. The results of each set of predictions can be used to calculate the numerical characteristics of clouds by Formulas (8)–(10) in Section 3.5 and then generating cloud droplets using Formula (11). Multiple cloud drops with numerical characteristics of clouds collectively compose the multidimensional cloud model.

Figure 15, below, shows the cloud model generated with the change in motion features at a certain moment. The vertical axis in the figure refers to the affiliation degree, which is a measure of the stability of the tendency of the current motion to belong to the target motion, and the horizontal axis is the feature values of the longitudinal features of the hand motion and the transverse features of the hand motion in the multidimensional cloud model. From the figure, it can be seen that the points with probabilities greater than 0.7 are more dense, indicating that most of the predictions are in accordance with the changes in real power operations when the STCNN prediction is performed.

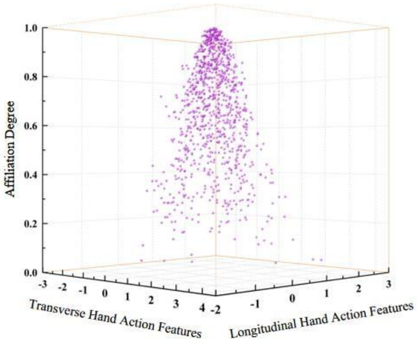


Figure 15. Cloud model generated by motion features at a certain moment.

Based on the multidimensional cloud model constructed by the actual action, the predicted action features are computed and generated as cloud drops and then compared in cloud space to obtain their probability of conforming to the change in power operation. All the collected frames in every 20 frames and all the previous frames can be used to build a 3D cloud model representing the actual action features. The 3D cloud model includes the offset features in x, y and z directions, and the affiliation of each cloud droplet in the set is actually a four-dimensional value.

As shown in Figure 16, below, the blue color indicates the multidimensional cloud model generated by the real motion features of the first 120 frames, and the green color indicates the multidimensional cloud model generated by the motion features of the next moment after the prediction of 110 frames. It can be seen that the two cloud models have a large difference, and the coordinate area with more dense points in green are the coordinate areas with affiliation degree scores lower than 0.7 in the blue cloud model; thus, it is decided that the action segmentation should be carried out at this time, the segmentation of the first action: turn to the tool table.

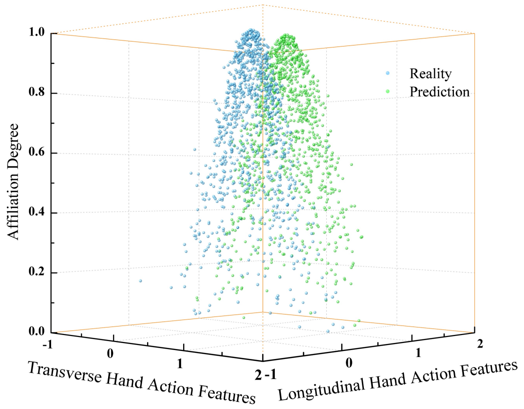


Figure 16. Comparison of cloud models for judgmental segmentation.

Based on the distribution of multi-outcome dataset predicted by STCNN in cloud space, the established multidimensional cloud model is a single-moment cloud model. When the probability value of the predicted operation action at a moment in the cloud model is lower than the expected probability, the moment is identified as the operation action segmentation point, thus completing the hand-action segmentation for continuous fine actions.

4.3. Experiments and Analysis of Results

In this paper, 50 sets of standard electrical detection operations are repeated, and the data are recorded and extracted to construct a hand spatial-temporal data matrix database to train and test the STCNN. A total of 80% of the data were used to train the model, and the remaining 20% of the data were used for testing; the experimental results of action feature-prediction accuracy are shown in Figure 17.

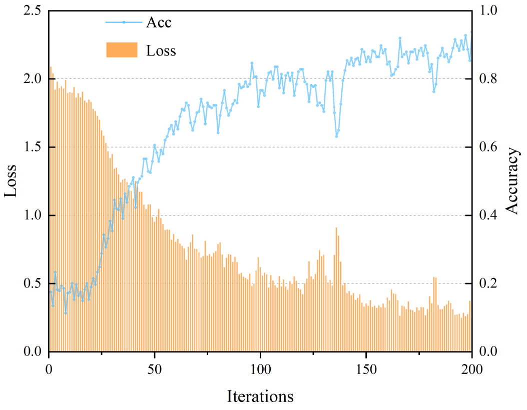


Figure 17. STCNN-model test results graph.

As seen in Figure 17, when the hand spatial-temporal data matrix is constructed with 10 frames, the accuracy of action feature prediction increases and the loss value decreases with the increase in training times, and the accuracy can reach 93.78% after 200 training times. In this paper, the prediction accuracy of both 2DCNN and BPNN were compared with that of STCNN after training 200 times, and the results are shown in Table 1. From the comparison results, it can be seen that the prediction accuracy of 2DCNN and BPNN is significantly lower than that of STCNN, because these two neural networks can only extract spatial features, which proves that STCNN can effectively predict the features of continuous fine actions.

Table 1. Comparison of prediction results of three methods.

Method	Epochs	Accuracy	Loss
STCNN	200	93.78%	0.2509
2DCNN	200	74.00%	0.6537
BPNN	200	63.76%	1.2119

According to the operation process of electrical inspection, the operation can be manually segmented into four phases, including moving, grasping, electrical testing and returning. The manual action segmentation combined with the workflow was compared with the action segmentation performed by the joint algorithm studied in this paper. The affiliation degree scores of the predicted operational action in the cloud model after using the joint algorithm for action segmentation is 0.512 at frame 94, 0.661 at frame 236, 0.663 at frame 367, 0.339 at frame 443, and 0.545 at frame 558, which is lower than the expected probability. The lower affiliation indicates the more obvious difference between the predicted action and the real action; therefore, five segmentation points are obtained and the overall operational action is divided into six segments. Combined with the whole operation process corresponding to the time, overall action segmentation included six sub-actions: steering tool table movement, grasp the electric pen, move to distribution box, contact test, return to the tool table, and put down the test pen. This is more than the four steps of the previous grid expert manual segmentation, but from the segmentation of the action results through the data, the segmentation of the action is more refined. The more points of division there are in the evaluation of continuous fine actions, the more details can be noticed, and the more detailed and comprehensive the evaluation of this action will be. The change in the distribution probability of action segmentation is shown in Figure 18, below.

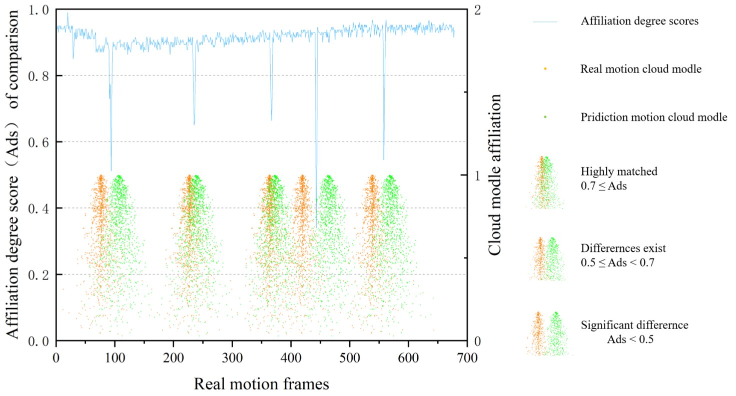


Figure 18. Segmentation results graph.

In this paper, the action segmentation of the operation process of electrical inspection using sliding-window and dynamic-programming methods is used as a comparison experiment, and the segmentation results of each method are shown in Table 2. The sliding-window and dynamic-planning methods segment the operation process of electrical inspection into five sub-actions, but the segmentation results are not as fine as the joint algorithm proposed in this paper, and there is a large discrepancy between the segmentation time and the real change time of the action. Neither the manual method nor the sliding-window and dynamic-programming methods segment the contact test action, which is an important scoring point in the overall electrical inspection training process.

Table 2. Comparison of different segmentation methods.

	Manual Segmentation	Sliding Window and Dynamic Programming	STCNN-MCM
Number of sub-actions	4	5	6
Result	Move to tool table. Electric pen pick up. Move to the electric box. Put back the electric pen.	Move to tool table. Electric pen pick up. Move to electric box. Return to the tool table. Put back the inspection pen.	Steer tool table movement. Grasp the electric pen. Move to distribution box. Contact test. Return to the tool table. Put down the test pen.

In this paper, the reliability and complexity of the procedure is measured by the number of sub-action segmentations and running time, and the STCNN-MCM joint algorithm outperforms the manual method and sliding-window and dynamic-programming methods in terms of the number of segmentations. In other vision-based action-segmentation methods, more time needs to be consumed for model training when the amount of video data is large [5]. The STCNN-MCM joint algorithm proposed in this paper achieves the effect of reducing the model complexity by introducing a cloud model in order to reduce the overall number of network layers and related parameters of the model. It took 252 s to train the STCNN-MCM on an NVIDIA GeForce RTX 1060 GPU. The runtime of the STCNN-MCM joint algorithm will outperform action segmentation using only a single neural network.

5. Discussion

Most of the current studies only perform action segmentation on common actions with distinct features in videos, but rarely on continuous fine actions in smart grid-training systems [33,34]. There are few studies on the segmentation of continuous fine actions during power operations. In this paper, STCNN-MCM is proposed for the action segmentation of power operations in smart grid training. Action segmentation with multidimensional cloud models can more significantly represent the differences between fine action features, making the determination results of action segmentation more objective and accurate.

Most of the current action-based training systems only complete the identification and segmentation of training actions based on spatial features or temporal features, which can analyze the standardization of actions and realize the correction and evaluation of trainers' actions [35,36]. In this paper, the completeness of feature extraction is improved by fusing the temporal and spatial information of hand-movement changes in the process of power operation to construct a hand spatial-temporal data matrix. In addition, the improved STCNN with added attention module is used to complete the prediction of the next moment's action feature, which contains both temporal and spatial information and has high prediction accuracy to correctly describe the fine action of the next moment.

In summary, this paper proposes a new action-segmentation algorithm based on the existing research, which can theoretically improve the segmentation efficiency of continuous fine actions in power operations. However, this algorithm requires highly accurate data extraction, and the features after combining the temporal and spatial information need to be highly correlated with the fine hand movements. At present, only the continuous fine action segmentation of the hands has been studied. In the subsequent process of improving the training of the smart grid, if the analysis of other parts of the body needs to be added, the importance of each part of the body in different actions needs to be further analyzed, and the combination of data will be carried out after determining the weights of each part. Limitations in data extraction and feature fusion will be improved in the subsequent work of the group.

6. Conclusions

In order to segment continuous fine actions during power operations, STCNN-MCM is proposed in this paper. Firstly, the temporal and spatial information of hand-movement changes during the power operation is fused to construct the hand spatial-temporal data matrix; then, the hand spatial-temporal data matrix is used as input, and the next-moment action features are predicted by the improved STCNN; finally, the corresponding probability is obtained based on the distribution of the predicted data in the cloud model, and the action segmentation point is determined based on this probability.

In application instances, this paper first validates the improved STCNN with an added attention module, which is trained based on the hand spatial-temporal data matrix dataset. From the test results, the improved STCNN can predict the motion features of fine hand actions in the next moment with high accuracy. Then, the operation process of electrical inspection in electric power operation as the experimental object, and the method proposed in this paper, were used for action segmentation. From the results of the segmentation, the segmentation of the action using this method is more precise. Therefore, STCNN-MCM can be applied to smart grid training to complete the segmentation of continuous fine actions in the process of power operation.

In the future, when using spatio-temporal convolutional neural networks to predict features for the next moment, a temporal attention module can be added to extract features of the action in the temporal dimension, so that the extracted features are more comprehensive and more consistent with the motion changes of the action. After the method is perfected, experiments on action segmentation will be conducted for other continuous fine actions during grid operations, to further improve the smart grid-training systems.

Author Contributions: Y.L. (Yong Liu): conceptualization, algorithm innovation, methodology, and writing—original draft; W.Z.: data and formal analysis, software, simulation, and writing—original draft; Y.L. (Yuan Li): formal analysis, investigation, and writing—review and editing; X.L.: conceptualization, simulation, investigation, methodology, and software; J.G.: investigation, visualization, and writing—review and editing; X.C.: supervision, and writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Mountaineering Management Center of the General Administration of Sport of China's critical project "Competitive Landscape and Strategies for Enhancing Competitive Strength in International Competitive Climbing" (CMA-2019-C06).

Data Availability Statement: Not applicable.

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

References

1. Aïm, F.; Lonjon, G.; Hannouche, D.; Nizard, R. Effectiveness of virtual reality training in orthopaedic surgery. *Arthroscopy* **2016**, *32*, 224–232. [CrossRef] [PubMed]
2. Çakiroğlu, Ü.; Gökoğlu, S. Development of fire safety behavioral skills via virtual reality. *Comput. Educ.* **2019**, *133*, 56–68. [CrossRef]
3. Butt, O.M.; Zulqarnain, M.; Butt, T.M. Recent advancement in smart grid technology: Future prospects in the electrical power network. *Ain Shams Eng. J.* **2021**, *12*, 687–695. [CrossRef]
4. Cikajlo, I.; Pogačnik, M. Movement analysis of pick-and-place virtual reality exergaming in patients with Parkinson's disease. *Technol. Health Care* **2020**, *28*, 391–402. [CrossRef] [PubMed]
5. Weinland, D.; Ronfard, R.; Boyer, E. A survey of vision-based methods for action representation, segmentation and recognition. *Comput. Vis. Image Underst.* **2011**, *115*, 224–241. [CrossRef]
6. Wu, D.; Shao, L. Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 724–731.
7. Guangming, Z.; Liang, Z.; Peiyi, S.; Juan, S. An Online Continuous Human Action Recognition Algorithm Based on the Kinect Sensor. *Sensors* **2016**, *16*, 161.
8. Lea, C.; Flynn, M.D.; Vidal, R.; Reiter, A.; Hager, G.D. Temporal convolutional networks for action segmentation and detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 156–165.

9. Kuehne, H.; Richard, A.; Gall, J. A hybrid rnn-hmm approach for weakly supervised temporal action segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *42*, 765–779. [CrossRef]
10. Singh, B.; Marks, T.K.; Jones, M.; Tuzel, O.; Shao, M. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1961–1970.
11. Farha, Y.A.; Gall, J. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3575–3584.
12. Wang, Z.; Gao, Z.; Wang, L.; Li, Z.; Wu, G. Boundary-aware cascade networks for temporal action segmentation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 34–51.
13. Gao, G.; Li, Z.; Huan, Z.; Chen, Y.; Liang, J.; Zhou, B.; Dong, C. Human behavior recognition model based on feature and classifier selection. *Sensors* **2021**, *21*, 7791. [CrossRef]
14. dos Santos, L.L.; Winkler, I.; Nascimento, E.G.S. RL-SSI Model: Adapting a Supervised Learning Approach to a Semi-Supervised Approach for Human Action Recognition. *Electronics* **2022**, *11*, 1471. [CrossRef]
15. Yao, G.; Lei, T.; Zhong, J. A review of convolutional-neural-network-based action recognition. *Pattern Recognit. Lett.* **2019**, *118*, 14–22. [CrossRef]
16. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 221–231. [CrossRef]
17. Molchanov, P.; Gupta, S.; Kim, K.; Kautz, J. Hand gesture recognition with 3D convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; pp. 1–7.
18. Lea, C.; Reiter, A.; Vidal, R.; Hager, G.D. Segmental spatiotemporal cnns for fine-grained action segmentation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 36–52.
19. Kopuklu, O.; Kose, N.; Gunduz, A.; Rigoll, G. Resource efficient 3d convolutional neural networks. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Republic of Korea, 27–28 October 2019.
20. Martin, P.E.; Benois-Pineau, J.; Péteri, R.; Zemmari, A.; Morlier, J. 3D Convolutional Networks for Action Recognition: Application to Sport Gesture Recognition. In *Multi-Faceted Deep Learning*; Springer: Berlin, Germany, 2021; pp. 199–229.
21. Chen, J.; Duan, H.; Song, Y.; Tang, M.; Cai, Z. CNN-Based Fluid Motion Estimation Using Correlation Coefficient and Multiscale Cost Volume. *Electronics* **2022**, *11*, 4159. [CrossRef]
22. Niu, Z.; Zhong, G.; Yu, H. A review on the attention mechanism of deep learning. *Neurocomputing* **2021**, *452*, 48–62. [CrossRef]
23. Kim, D.H.; Anvarov, F.; Lee, J.M.; Song, B.C. Metric-based attention feature learning for video action recognition. *IEEE Access* **2021**, *9*, 39218–39228. [CrossRef]
24. Zhu, M.; Zhu, X.; Zhu, C. STGATP: A Spatio-Temporal Graph Attention Network for Long-Term Traffic Prediction. In Proceedings of the International Conference on Artificial Neural Networks, Bristol, UK, 6–9 September 2021; pp. 255–266.
25. He, F.; Liu, Y.; Zhan, W.; Xu, Q.; Chen, X. Manual Operation Evaluation Based on Vectorized Spatio-Temporal Graph Convolutional for Virtual Reality Training in Smart Grid. *Energies* **2022**, *15*, 2071. [CrossRef]
26. Qiao, B.; Wu, Z.; Tang, Z.; Wu, G. Sea surface temperature prediction approach based on 3D CNN and LSTM with attention mechanism. In Proceedings of the 2022 24th International Conference on Advanced Communication Technology (ICACT), Phoenix Pyeongchang, Republic of Korea, 13–16 February 2022; pp. 342–347.
27. Wei, L.; Guo, D.; Chen, Z.; Yang, J.; Feng, T. Forecasting Short-Term Passenger Flow of Subway Stations Based on the Temporal Pattern Attention Mechanism and the Long Short-Term Memory Network. *ISPRS Int. J. Geo-Inf.* **2023**, *12*, 25. [CrossRef]
28. Li, D.; Liu, C.; Gan, W. A new cognitive model: Cloud model. *Int. J. Intell. Syst.* **2009**, *24*, 357–375. [CrossRef]
29. Deng, Y.; Liu, S.; Zhang, W.; Wang, L.; Wang, J. General multidimensional cloud model and its application on spatial clustering in Zhanjiang, Guangdong. *J. Geogr. Sci.* **2010**, *20*, 787–798. [CrossRef]
30. Yao, J.; Wang, G.; Xue, B.; Wang, P.; Hao, F.; Xie, G.; Peng, Y. Assessment of lake eutrophication using a novel multidimensional similarity cloud model. *J. Environ. Manag.* **2019**, *248*, 109259. [CrossRef]
31. Wang, M.; Liu, Q.; Wang, X.; Shen, F.; Jin, J. Prediction of rockburst based on multidimensional connection cloud model and set pair analysis. *Int. J. Geomech.* **2020**, *20*, 04019147. [CrossRef]
32. Zhang, R.L.; Liu, X.H. A Novel Hybrid High-Dimensional PSO Clustering Algorithm Based on the Cloud Model and Entropy. *Appl. Sci.* **2023**, *13*, 1246. [CrossRef]
33. Ahn, H.; Lee, D. Refining action segmentation with hierarchical video representations. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 11–17 October 2021; pp. 16302–16310.
34. Yi, F.; Wen, H.; Jiang, T. Asformer: Transformer for action segmentation. *arXiv* **2021**, arXiv:2110.08568.
35. Tsai, W.L.; Su, L.w.; Ko, T.Y.; Yang, C.T.; Hu, M.C. Improve the decision-making skill of basketball players by an action-aware VR training system. In Proceedings of the 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Osaka, Japan, 23–27 March 2019; pp. 1193–1194.
36. Xin, L. Evaluation of factors affecting dance training effects based on reinforcement learning. *Neural Comput. Appl.* **2022**, *34*, 6773–6785. [CrossRef]

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.