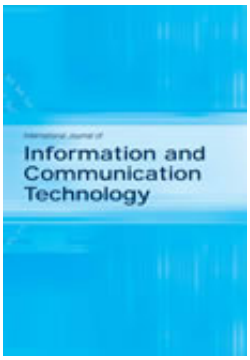
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Ethnic dance movement recognition based on motion capture sensor and machine learning

Mengying Li

Article History:

Received:	30 August 2024
Last revised:	14 September 2024
Accepted:	16 September 2024
Published online:	17 October 2024

Ethnic dance movement recognition based on motion capture sensor and machine learning

Mengying Li

School of Music,
Suihua University,
Suihua 152000, China
Email: limengyingshxy@163.com

Abstract: To address the shortcomings of the existing folk dance movement recognition techniques in terms of accuracy, real-time and generalisation ability, this paper proposes an innovative research method, i.e., combining the motion capture technology with the machine learning algorithm 3D convolutional neural network. This paper describes in detail the various aspects of the research method, including the acquisition of motion data, preprocessing, feature extraction and selection, and the construction and training of machine learning models. Meanwhile, we adopt a variety of high-precision sensors to accurately capture the details of the dancer's movements, and conducts in-depth learning and analysis of the movement features by the 3D convolutional neural network in machine learning. Finally, the experimental results show that the method proposed significantly exceeds the traditional method in terms of the accuracy, robustness, and real-time performance of folk-dance movement recognition, which proves the effectiveness and superiority.

Keywords: motion capture; machine learning; folk dance; motion recognition; deep learning.

Reference to this paper should be made as follows: Li, M. (2024) 'Ethnic dance movement recognition based on motion capture sensor and machine learning', *Int. J. Information and Communication Technology*, Vol. 25, No. 8, pp.81–96.

Biographical notes: Mengying Li is currently a Teacher of Dance Choreography at Suihua University. Her research interests include choreography, special education dance, and multimedia dance.

1 Introduction

In the wave of the digital era, folk dance recognition technology is gradually becoming a bridge connecting traditional culture and modern technology. Folk dance, with its unique style and expression, not only plays an important role in cultural inheritance, but also provides rich materials for modern artistic creation and performance (Xie, 2023). However, the automatic recognition of folk dances faces many challenges, including the high dimensionality and diversity of movements as well as subtle emotional expressions, which make the development of an automated recognition system complex and challenging. Therefore, the development of an efficient and accurate automated folk

dance movement recognition technology not only has important academic value, but also has a wide range of application potential.

It is in this context that the combination of motion capture sensor technology and machine learning techniques provides new ideas to address these challenges. Motion capture sensor technology, as a breakthrough in modern computer graphics and human-computer interaction, has significantly changed the way we capture and analyse human movement. Through a series of high-precision sensors, this technology is able to monitor and record in real time the tiny movements of the dancer's body parts, transforming the movements into quantifiable data (Shi, 2022). These sensors capture the dancer's precise position and posture in space, including multi-dimensional information such as speed, direction and acceleration, providing an unprecedentedly detailed view of the digital recording of folk dance. At the same time, the development of machine learning technologies, especially in the field of deep learning, provides advanced algorithmic support for analysing these complex datasets. Deep learning models, with their powerful data processing and automatic feature extraction capabilities, can learn the intrinsic patterns and features of dance movements from motion capture data (Idris et al., 2017). These models, through training, are able to recognise and classify different dance movements and even understand the nuances and emotional expressions between dance movements.

With the development of technology, there are numerous domestic and international researches on motion capture, and motion capture technology is widely used in human movement. The Kinect V2 depth sensor launched by Microsoft has advanced motion capture technology, which is able to collect three kinds of data: RGB map, depth map and joint points of human skeleton (Yang et al., 2019). Therefore, in recent years, there have been numerous researches combining the Kinect depth sensor with motion recognition in the fields of medical treatment, public monitoring, automatic driving, entertainment, sports and cultural digital preservation, and so on. Hu et al. (2021) used the human bone tracking technology in Kinect in dance assisted training, and proposed a representation method of skeletal joint point angles based on fixed axes, which improves the stability of the data during the measurement of the joint point angles, and improved the human posture recognition method based on the angle of the joints on the basis of this method, and developed a dance assisted training system based on Kinect. Kitsikidis et al. (2015) used multiple Kinect sensors to capture dance movements for the first time in order to solve the occlusion and self-occlusion tracking problems. The fused skeletal data was divided into five different body parts and then transformed to allow view-invariant pose recognition, demonstrating the high recognition accuracy of the proposed method. Protopapadakis et al. (2017) also used Kinect sensors to capture six Greek folk dance movements and compared the classification results using four commonly used classifiers to classify the movements directly on the raw data. The effect of different human joints on the recognition rate was also investigated.

Machine learning techniques play a central role in the exploration of folk dance movement recognition, especially the application of 3D convolutional neural networks. Kamnitsas et al. (2017) first proposed 3D CNNs architecture for movement recognition. 3D convolutional networks are a direct extension of 2D convolutional networks, and 3D convolutional networks have one more dimension of capture time information than 2D convolutional networks. Ji et al. (2013) proposed a 3D CNNs architecture, which generates information from adjacent video frames in multiple channels and performs convolution and subsampling in each channel separately, and then obtains the final

feature representations by synthesising the information from each channel. Bargellesi et al. (2019) based on the former proposed a modern deep architecture of C3D (convolutional 3D) based on the former, which can be learned on large-scale datasets. Xu et al. (2018a) proposed an asymmetric three-dimensional convolutional neural network (3D-CNN) method for action recognition task, which is able to minimise the need to train two networks on RGB and optical flow fields training the two networks separately, which improves the computational efficiency. Duan et al. (2022) proposed PoseC3D, a 3D CNN-based skeleton recognition method, which can extract spatio-temporal features in human skeleton sequences more efficiently, is more robust to noise in skeleton sequences, and has better generalisation. These research results not only demonstrate the effectiveness and potential of motion capture and machine learning techniques in folk dance movement recognition, but also provide rich experience and data support for the development and application of related technologies in the future, and the accuracy and utility of folk dance movement recognition techniques will be improved even more. As a key technology for dance information retrieval and cultural inheritance, folk dance movement recognition plays a crucial role in building an intelligent dance teaching system, enhancing the interactive experience of folk dance, and promoting the innovative development of dance art. Although existing research has made some progress in folk dance movement recognition, the existing technology still faces a number of challenges when dealing with complex and changing folk dance movement data. In particular, there are obvious limitations of traditional methods in terms of the accurate capture of movement features and the ability of models to generalise to diverse dance styles. To address these challenges, this study proposes an innovative solution that combines motion capture techniques and machine learning algorithms to overcome the limitations of existing techniques. With this approach, we are not only able to automatically extract the spatio-temporal features of folk-dance movements, but also significantly improve the model's recognition accuracy and adaptability to different dance movements.

The contributions of the thesis are mainly in the following areas:

1 This study innovatively applies 3D convolutional neural networks to folk dance action recognition, which not only enhances the model's ability to capture spatio-temporal features of the action, but also improves the accuracy of the recognition. Compared with the traditional two-dimensional convolutional neural network, the three-dimensional network can process the video data more efficiently and capture the time-series features of the dance movements, thus realising more accurate movement recognition.

2 By using a variety of high-precision sensors, this study realises the accurate capture of dancer movement details, and the fusion of these sensor data provides a more comprehensive and detailed input for movement recognition. This data fusion technique significantly improves the accuracy and robustness of motion capture, and lays a solid foundation for subsequent feature extraction and analysis.

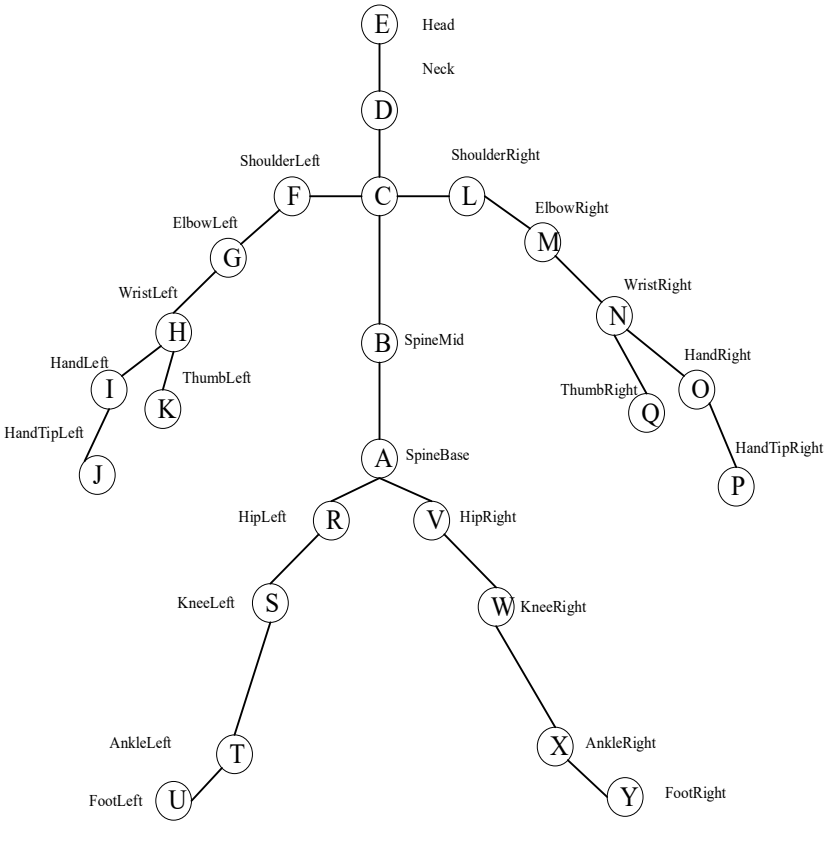
3 In this paper, we propose a complete end-to-end process covering action data acquisition, preprocessing, feature extraction and selection, up to the construction and training of machine learning models. The innovation of this process lies in its systematic and integrated nature, which ensures that every step from the raw data to the final recognition result is optimised, improving the efficiency and accuracy of the whole recognition system.

2 Relevant theoretical analysis

2.1 Kinect sensor

The Kinect sensor, developed by Microsoft, functions as a sophisticated motion-sensing apparatus, equipped with a suite of features including a high-definition colour camera, infrared transmission capabilities, and a depth-sensing camera, all of which facilitate real-time motion tracking and skeletal recognition, complemented by audio capture (Chen and Koskela, 2015). This has solidified the Kinect as a pivotal instrument in the domain of motion capture research. The paper utilises the Kinect v2.0, an enhanced successor released by Microsoft in June 2014. This upgraded model showcases significant performance improvements over the initial Kinect, offering refined motion detection and analysis capabilities (Lachat et al., 2015). Kinect V2 can be captured in real time to the human body's 25 skeleton joints, and each of these joints is tracked, inferred, and tracked. Each skeletal joint point has three states: tracked, inferred, and not tracked. Skeletal points are connected by line segments, and the three-dimensional spatial information of the human body can be determined according to the establishment of the coordinate system, so that the relevant human skeleton model can be captured. Figure 1 shows the human skeletal joint point model under Kinect V2.

Figure 1 Human skeletal joint point model



The skeletal joint point model provides a detailed representation of a dancer's movements by tracking specific joints. This precision allows for accurate capture of complex dance movements, essential for effective recognition. Currently, positional information such as skeletal 3D coordinate points and poses of the human body can be represented using the relevant API functions in the Kinect SDK, while Kinect V2 can recognise 25 human skeletal point information. When performing the matching between the captured action and the database action joint points, the nodes of the reproduction model are different due to the different force of each joint point. Consequently, to enhance the alignment accuracy of joint points and minimise the model's reconstruction discrepancies, this study adopts the hip joint of the dance trainer as a reference anchor in spatial metrics. Utilising this reference, the spatial disparity for each joint point is quantified, delineated by the subsequent equations [equations (1) and (2)]:

$$d_i = \sqrt{\sum_{j=1}^{N} \left(\Omega_i^j - \Omega_\tau^j\right)^2 / \sigma_j} \tag{1}$$

$$d_j = \sum_{i=1}^{N} \left(M - j_i\right)^2 / N \tag{2}$$

where σ_j denotes the standard deviation of each joint point in the model under the overall action; M and j_i denote the position of joint point j in the action and the position of frame i in the human action data, respectively; Ω_i^j and Ω_τ^j denote the key position of joint point j under frame i and frame τ, respectively.

The Kinect device harnesses the principle of stereo triangulation to measure the depth of objects within its field of view. An object, denoted as P, is situated at a three-dimensional coordinate (X, Y, Z) in the global reference system. The term 'baseline' refers to the spatial gap, B, between the epicentres of the dual lenses, while the XZ plane corresponds to the level where the optical planes intersect, aligning the X-axis with the baseline and positioning the Y-axis orthogonal to the optical axis. The depth, Z, is ascertained by leveraging the parallax observed between the two cameras, assuming their optical axes are parallel – a technique conventionally termed as triangulation. This process is mathematically articulated through equations (3), (4) and (5).

$$Z = \frac{(B \times d)}{(X_2 - X_1)} \tag{3}$$

$$X = X_1 \times Z / d \tag{4}$$

$$Y = Y_1 \times Z / d \tag{5}$$

Identifying the body parts and joint points requires the use of a classifier that contains many depth-informed features as shown in equation (6), using which the body parts can be determined.

$$f_\theta(I, x) = d_I\left[x + \frac{u}{d_I(x)}\right] - d_I\left[x + \frac{v}{d_I(x)}\right] \tag{6}$$

where x represents the pixel value at each point, $d_I(x)$ represents the depth value corresponding to the pixel value at point x in Figure 1, $\theta = (u, v)$ is a parameter that contains u and v, both representing offset vectors, and $1/d_I(x)$ is an offset regularisation representing the difference between the u and v depth offsets. These features are related to the 3D shape, allowing them to be used as features for machine learning classifiers capable of recognising body parts and joint points.

2.2 Convolutional neural networks

Convolutional neural networks (CNN) have their origins in multilayer perceptrons, and their development can be traced back to Yann LeCun's LeNet-5 model inspired by the cat's visual cortex. In 2012, AlexNet's breakthrough victory in the ImageNet competition marked the rise of CNNs. Conventional convolutional neural networks are typically built up from an input layer, a convolutional layer, a pooling layer, a fully-connected layer, and an output layers built by connecting them in a hierarchical manner (Duan et al., 2022).

A convolutional layer is instrumental in the feature extraction process of input data, encompassing an array of convolutional kernels, alternatively referred to as filters. These kernels traverse the input matrix, detecting patterns and creating feature maps that capture the underlying structure within the data. Commonly, the convolutional kernel is a feature extractor, each convolutional kernel generally corresponds to a class of features, such as the vertical texture in the picture, and each element in a convolutional kernel has a weight and a bias value (Yao et al., 2019). When working, the convolution kernel moves according to a predetermined step (stride), performs matrix dot product operations on the swept region, and superimposes the bias values, and the mathematical expression of the convolution layer is shown in equation (7):

$$x_j^l = f\left(\sum_{i \in M_j} x_i^{l-1} * \omega_{ij}^l + b_j^l\right) \tag{7}$$

where ω_{ij}^l and b_j^l denote the weight and bias corresponding to the convolutional filter at position (i, j), respectively; x_i^{l-1} denotes the feature mapping of the previous layer; x_j^l denotes the feature mapping of the current layer; and M_j denotes the set of feature mappings.

The outputs from convolutional layers, known as feature maps, are typically directed to a pooling layer for the purpose of feature selection and dimensionality reduction through a process termed downsampling. This operation is parameter-free; thus, it does not contribute to the model's training parameters but is instrumental in mitigating overfitting. The pooling operation shares similarities with the convolutional kernel scanning process; its output dimensions are determined by factors such as the size of the pooling area, the stride, and the padding. Prevalent pooling techniques include maximum pooling, which selects the largest value within the pooling window, mean pooling, which averages the values, and stochastic pooling, which randomly samples a subset of the window. To introduce nonlinearity and enhance the model's capacity for complex mappings, an activation function is often interposed between the convolutional and pooling layers. Standard activation functions include sigmoid, tanh, and ReLU. The sigmoid function, for instance, produces outputs within the interval [0, 1], effectively normalising the neuron outputs, and is defined by equation (8).

$$sigmoid(x) = \frac{1}{1+e^{-x}} \tag{8}$$

The output value of tanh is in the range of [1, 1], specifically, the output value is 1 when the positive number is larger and –1 when the negative number is larger. The calculation of tanh is shown in equation (9):

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{9}$$

When the input value is less than 0, the output value of ReLU is 0 and the derivative value is also 0. This will lead to the neuron not being able to perform parameter update and the phenomenon of gradient vanishing. The calculation of ReLU is shown in equation (10):

$$ReLU(x) = \max(0, x) \tag{10}$$

Currently, within the architecture of neural networks, average pooling and maximum pooling are prevalent techniques employed for the pooling operations. These methods are commonly utilised to reduce the spatial dimensions of the feature maps, thereby facilitating the construction of efficient network models. According to equation (11), the average pooling operation takes the average value corresponding to the pooled region; the maximum pooling operation takes the maximum value corresponding to the pooled region, and its mathematical expression is shown in equation (12).

$$pooling_{u,v}^{\max} = \frac{1}{\left|\Omega_{u,v}\right|} \sum\nolimits_{i,j \in \Omega_{u,v}} a_{i,j} \tag{11}$$

$$pooling_{u,v}^{average} = \max_{i,j \in \Omega_{u,v}} a_{i,j} i, \quad j \in \Omega_{u,v} \tag{12}$$

where $a_{i,j}$ is the activation value of the pooled region; i, j are the index representations; $\Omega_{u,v}$ is the corresponding pooled region on the feature map.

The fully connected layer in a convolutional neural network is generally two or three layers connected together and placed before the output layer, the main role is to tile the high-dimensional multi-channel data into one-dimensional vectors to facilitate subsequent calculations, also known as the dense layer in some deep learning frameworks.

The last layer in the output layer convolutional neural network. In the recognition task, the number of neurons in this layer corresponds to the ethnic dance recognition and is followed by a softmax activation function to compute the discriminative probability distribution of the input image in each category. After the softmax activation function processing maps the network model's scores for the input data into the (0, 1) interval, the output of the softmax activation function is the network model's discriminative probability for the input samples in each category. The expression of the softmax activation function is equation (13):

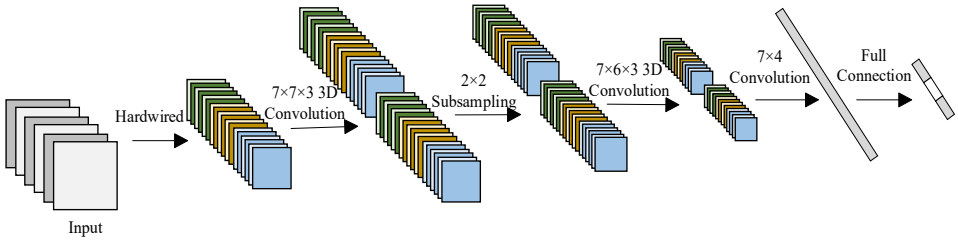
$$a_i = \frac{e^{z_i}}{\sum_{k=1}^{m} e^{z_k}} \tag{13}$$

where z_i is the score of the network model for category i and a_i is the predicted probability value of the input on category i. The maximum probability value of determines the category of the result and the sum of all is one. The maximum probability value of a_i determines the category of the prediction result and the sum of all a_i is 1. The network model can calculate the loss value for this training by using the probability distribution of the output and the true labels of the input samples during the training process.

2.3 Three-dimensional convolutional neural networks

In contrast to conventional deep learning approaches, 3D CNNs transcend the limitation of processing solely 2D single-frame images. They possess the capability to capture features from both the spatial and temporal domains, enabling the extraction of motion cues across sequential frames. This paper leverages 3D CNNs to discern the skeletal data of quintessential folk dance movements, which, being single-channel, offers reduced computational demands and enhanced model recognition efficacy (Shotton et al., 2013). The framework of the 3D CNNs model employed in this study is depicted in Figure 2. The model comprises four convolutional layers, interspersed with two max-pooling downsampling layers that utilise a 3 × 3 × 3 kernel, followed by two fully connected layers, culminating in a softmax layer dedicated to classification. This architectural design is tailored to optimise the feature extraction and classification accuracy for the task at hand.

Figure 2 3D convolutional neural network framework (see online version for colours)



To effectively seize the dynamic elements across a sequence of frames, the feature computation encompasses both spatial and temporal extents. The formulation for the value within the j^{th} feature map of the i^{th} layer, at a specific cell indexed by the coordinates (x, y, z), is articulated in equation (14) as depicted below.

$$V_{ij}^{xyz} = f\left(b_{ij} + \sum_{r}\sum_{l=0}^{l_i-1}\sum_{m=0}^{m_i-1}\sum_{n=0}^{n_i-1}\omega_{ijr}^{lmn} v_{(i-1)r}^{(x+l)(y+m)(z+n)}\right) \tag{14}$$

where the time dimension of the 3D convolutional kernel is n_i and the weight value of the convolutional kernel for the location (l, m, n) connected to the r^{th} feature map is ω_{ijr}^{lmn}.

The ReLU serves as a prevalent activation function within deep learning architectures. It maintains the input feature value in its original form for outputs where the value exceeds zero, while mapping negative inputs to zero. This characteristic introduces a form of thresholding that promotes sparsity in model parameters, mitigating

the likelihood of overfitting (Xing and Zhu, 2021). Moreover, the simplicity of the ReLU derivative calculation aids in accelerating the training process. As the derivative of the ReLU function is consistently 1 for positive inputs, it effectively combats the vanishing gradient issue. The mathematical expression for the ReLU activation function is given by equation (15).

$$f(x) = \max(0, x) = \begin{cases} 0, & x \le 0 \\ x, & x > 0 \end{cases} \tag{15}$$

Pooling, alternatively termed downsampling, involves a unique consideration for video data along the temporal axis beyond traditional 2D image processing. This operation effectively condenses the feature map, thereby diminishing the data's spatial and volumetric extent, which in turn lessens computational load and facilitates more tractable training, potentially enhancing model accuracy (Song et al., 2018). The process of maximum pooling is articulated through the formulation presented in equation (16).

$$V_{x,y,z} = \max_{0 \le i \le s_1, 0 \le j \le s_2, 0 \le k \le s_3} \mu_{x \times s+i, y \times t+j, z \times r+k} \tag{16}$$

where μ symbolises the input vector within a three-dimensional space, while V represents the resultant output post-pooling operation, with s, t and r indicating the sampling intervals along their respective dimensions. The softmax function, commonly deployed in the final layer of classification models, transforms an n-dimensional input vector x into a probability distribution. This transformation ensures that the probability of the correct class approaches 1, the probabilities of incorrect classes approach 0, and the sum of probabilities across all classes equals unity.

3 Ethnic dance movement recognition based on motion capture sensor and machine learning

3.1 Folk dance characteristics

The typical movement dataset of minority dances includes five types of folk dances, each type of dance contains four different movements, totalling 20 movements. We present RGB graphs of selected minority dance movement sequences, and each subsequent movement will be represented by a set of consecutive skeleton sequences. The ethnic minority dance categories studied include:

1 Dai dance: the Dai dance, emblematic of the Dai ethnicity in the southwestern regions, predominantly draws its choreography from mimicking the natural behaviours of indigenous fauna. Characterised by a signature pose known as the 'three bends', this dance form involves dancers gracefully inclining their upper bodies to one side while partially squatting, creating an inverted 'S' silhouette across their head, chest, waist, and limbs. The Dai dance is distinguished by its fluidity and minimal reliance on leaping motions.

Figure 3 Typical movements of Dai dance (see online version for colours)

2 Tibetan dance: Tibetan dance, a hallmark of Tibetan cultural expression, is deeply rooted in the highland terrain. The dance's vigour is predominantly channelled through the lower body, with dancers articulating an aesthetic appeal through the rhythmic bending and straightening of their joints. In Tibetan dance, the body's focal point leans forward, with arms often left to dangle naturally. Female dancers exhibit an air of elegance and poise, while their male counterparts project strength and robustness. The dance encapsulates the artistic soul of the Tibetan people, offering a glimpse into their rich historical tapestry.

Figure 4 Typical movements of Tibetan dance (see online version for colours)

3 Viennese dance: the Viennese dance has the stylised sense of standing upright, with movements characterised by the head, neck, shoulders, chest, waist and feet. The Wei dance is distinguished by its posture, characterised by an upright stance that elevates the head and straightens the chest, conveying an impression of nobility and pride, as well as an open and upright demeanour. The movements of the Wei dance are beautifully styled and changeable, and with the dancers' eyes, neck movement, finger snapping and finger snapping, etc., it shows the enthusiasm and joyfulness of the Wei dance.

Figure 5 Typical movements of Uyghur dance (see online version for colours)

4 Mongolian dance: Mongolian dance is characterised by a large range of movements and a fast rhythm. Key to the Mongolian dance are expressive shoulder shrugs, wrist rotations, and fluid arm gestures. Dancers are expected to embody a warm, courageous, and free-spirited persona, with hand movements that are supple and in sync with the music's cadence, thereby showcasing the bravery characteristic of the Mongolian people. The dance also serves as a medium to convey the vast prairie landscapes, cultural practices, and the essence of the Mongolian people.

Figure 6 Typical movements of Mongolian dance (see online version for colours)

5 Miao dance: in this paper, we focus on the Miao Jinji dance, a cultural heritage from the Qiandongnan area of Guizhou Province. The Jinji, revered as a totemic symbol by the Miao, is celebrated and used as a medium to honor their ancestors through this dance. Performers adhere to the rhythm of the Lusheng, an indigenous instrument, moving in an anticlockwise manner. They transition between movements in time

with varying beats, maintaining a natural swing of the upper limbs and feet while keeping their knees in a slight bend, embodying the dance's traditional essence.

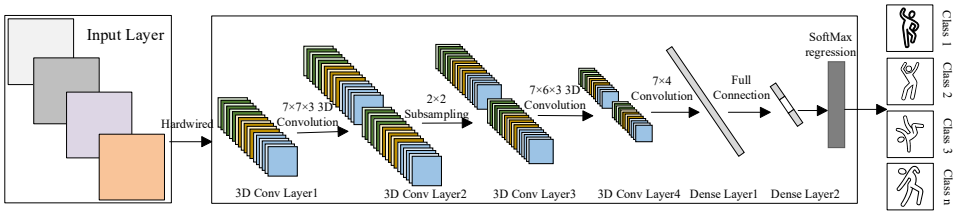
Figure 7 Typical movements of Miao dance (see online version for colours)

3.2 Ethnic dance recognition model

This paper presents an innovative folk dance movement recognition framework that integrates motion capture techniques and machine learning algorithms, as shown in Figure 8. The methodological basis adopted in this study lies in the fact that, at the initial stage, a publicly available folk-dance dataset is used to train a 3D-CNN model aimed at accurately tuning the network weights so that the network is able to capture and characterise complex dance movement features. Further, we apply this pre-trained network to a new scenario of ethnic dance movement recognition, utilising it as a feature extraction tool to deeply mine dance data captured from sensors.

The folk dance action recognition process of this method is carefully designed into four main sessions: firstly, the pre-training session of 3D-CNN model, which is trained with rich dance action samples; followed by the feature extraction session, in which the pre-trained model is used to extract key features from the action data; and then the classifier training session, in which the classifier is trained based on the extracted features to achieve accurate recognition of diverse folk dance movements; and finally the movement category probability prediction session, in which unknown dance movement samples are recognised and their categories are predicted (Xu et al., 2018b). This series of organised steps ensures the efficiency and accuracy of this method in automated folk dance movement classification and recognition.

Figure 8 Model framework diagram (see online version for colours)



4 Experiments and results

This study has amassed a dataset, ETHDance, comprising a collection of 600 instances of ethnic dance movement skeleton data. It encompasses five distinct ethnic dance forms, with each form featuring four distinct motion categories, culminating in a total of 20 unique motion categories. For each specific motion, 30 samples have been meticulously gathered. The labelling of the movements is reflected by a specific file naming, the file is named in the format of a00_s00_e00, where a indicates the movement serial number, s indicates the character subject, and e indicates the number of times the movement was performed, and a01_s01_e01, for example, indicates the movement data of the 1st person performing the 1st movement for the 1st time.

The experiments for this study were conducted extensively using several public datasets, including the UTKinect dataset, MSRAction3D dataset, and the in-house ethnic dance movement dataset. Both the UTKinect and MSRAction3D datasets were captured using Kinect depth sensors, providing a comprehensive set of 20 joint movement data points that align with the structure of the dance movement data presented in this paper. The UTKinect dataset features 10 action categories with 200 samples, while the MSRAction3D offers a broader range with 20 action categories and 567 samples. Our dataset, ETHDance, comprises 20 categories of dance actions with 600 samples in total. The methodology presented in this paper achieved an 81% recognition rate on the UTKinect dataset, 91% on the MSRAction3D dataset, and an impressive 95% on the ETHDance dataset. These results not only substantiate the efficacy of our approach across various action recognition datasets but also reflect the robustness and rationality of the dataset we have curated. The consistent high performance across different datasets underscores the generalisability and reliability of the proposed 3D CNNs model.

Table 1 Model comparison results

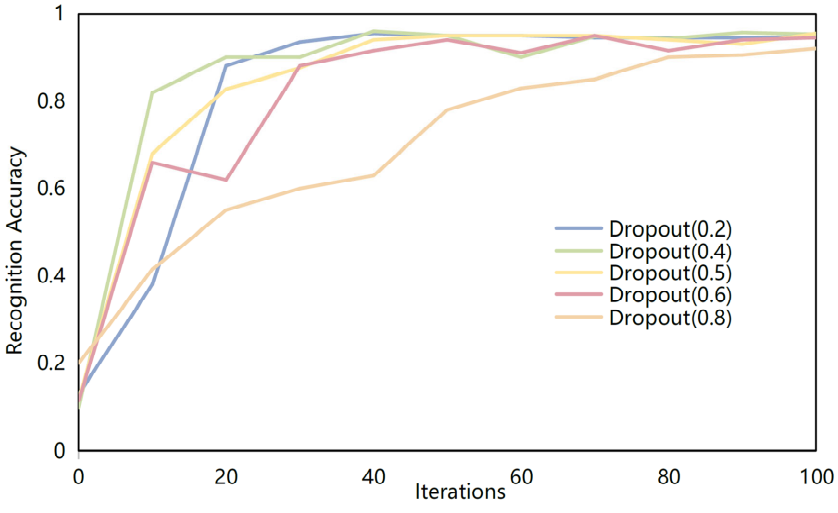
Classification model	*ACC*
3D-CNN-JM	88.33%
Shallow 3D CNN	93.33%
Efficient 3D CNN	91.67%
Our 3D-CNN	95%

To substantiate the efficacy of the 3D CNNs approach outlined in this paper, comparative experiments were conducted with other established algorithms on the ethnic dance movement dataset presented here, including 3D-CNN-JM (Wang et al., 2019), shallow 3D CNN (Singh et al., 2020) and efficient 3D CNN (Wang et al., 2023). Our-3D-CNN has an accuracy of 95%, while shallow-3D-CNN, efficient-3D-CNN and 3D-CNN-JM have accuracy rates of 93.33%, 91.67% and 88.33%, respectively. Therefore, compared with shallow-3D-CNN, the accuracy of our-3D-CNN is improved by 1.67%; compared with efficient-3D-CNN, our-3D-CNN is improved by 1.67%; compared with efficient-3D-CNN, efficient-3D-CNN is improved by 1.67%. Our-3D-CNN improves the accuracy by 3.33% compared to efficient-3D-CNN, and our-3D-CNN improves the accuracy by 1.67% compared to 3D-CNN-JM. CNN improves the accuracy by 6.67%. As detailed in Table 1, a comparative analysis of the experimental outcomes from other methodologies within the scope of this paper's ethnic dance movement dataset reveals that they too exhibit commendable recognition

capabilities. These findings collectively validate the sound construction of our dataset and the robustness of the recognition model proposed in this study.

Furthermore, to mitigate the issue of overfitting, this paper employs the dropout technique with varying ratios. Experiments were conducted using ratios of 0.2, 0.4, 0.5, 0.6, and 0.8 on the dataset presented in this paper, with validation results depicted in Figure 9. Each condition was iterated 100 times to assess the impact of different Dropout ratios on test set recognition accuracy. The results indicated that the test set accuracy was marginally higher with dropout ratios of 0.4 and 0.5, particularly after 50 iterations. Among these, a dropout ratio of 0.5 yielded a slightly better recognition accuracy compared to the 0.4 ratio. Consequently, all subsequent experiments utilised a dropout ratio of 0.5, which demonstrated optimal performance in terms of recognition accuracy.

Figure 9 Dropout ratio experiment (see online version for colours)



5 Conclusions

In this study, we successfully constructed an end-to-end process from data acquisition to recognition by combining advanced motion capture technology with machine learning algorithms, providing a new approach for automatic recognition of folk dance movements. Using the ethnic dance skeleton data collected by the Kinect depth sensor, we carefully curated a dataset that is free from background and illumination interference, and processed it by our optimised 3D CNNs to achieve accurate recognition of dance movements. Experimental results show that our method exhibits excellent recognition performance on different datasets, validating the effectiveness and generalisability of our method. We draw the following conclusions:

1 Effectiveness of technology integration: this study demonstrates that the effective combination of motion capture technology and 3D CNNs in folk dance movement recognition can accurately capture the spatio-temporal features of the dance and significantly improve the accuracy and robustness of the recognition.

2 Importance of the dataset: the construction of the ETHDance dataset provides rich samples of folk dance movements for this study, and its high quality and diversity are the key factors to achieve high recognition rates.

3 Superiority of the model: through comparison experiments with other algorithms, our 3D CNNs model demonstrates higher recognition accuracy on the folk dance movement dataset, proving the superiority of the model design.

4 Algorithm's anti overfitting ability: the introduction of dropout technology effectively improves the model's generalisation ability, and the experimental results show that an appropriate dropout ratio can further improve the model's recognition accuracy.

References

Bargellesi, N., Carletti, M., Cenedese, A. et al. (2019) 'A random forest-based approach for hand gesture recognition with wireless wearable motion capture sensors', *IFAC-PapersOnLine*, Vol. 52, No. 11, pp.128–133.

Chen, X. and Koskela, M. (2015) 'Skeleton-based action recognition with extreme learning machines', *Neurocomputing*, Vol. 149, pp.387–396.

Duan, H., Zhao, Y., Chen, K. et al. (2021) 'Revisiting skeleton-based action recognition', *Conference Revisiting Skeleton-based Action Recognition*, IEEE.

Hu, H., Cao, Z., Yang, X. et al. (2021) 'Performance evaluation of optical motion capture sensors for assembly motion capturing', *IEEE Access*, Vol. 9, pp.61444–61454.

Idris, M.Z., Mustaffa, N., Othman, A.N. et al. (2017) 'Exploring principle components for digital heritage preservation on Malay folk dances', *International Journal of Academic Research in Business and Social Sciences*, Vol. 7, No. 10.

Ji, S., Xu, W., Yang, M. et al. (2013) '3D Convolutional neural networks for human action recognition', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 1, pp.221–231.

Kamnitsas, K., Ledig, C., Newcombe, V.F.J. et al. (2017) 'Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation', *Medical Image Analysis*, Vol. 36, pp.61–78.

Kitsikidis, A., Boulgouris, N.V., Dimitropoulos, K. et al. (2015) 'Unsupervised dance motion patterns classification from fused skeletal data using exemplar-based HMMs', *International Journal of Heritage in the Digital Era*, Vol. 4, No. 2, pp.209–220.

Lachat, E., Macher, H., Landes, T. et al. (2015) 'Assessment and calibration of a RGB-D camera (Kinect v2 sensor) towards a potential use for close-range 3D modeling', *Remote Sensing*, Vol. 7, No. 10, pp.13070–13097.

Protopapadakis, E., Grammatikopoulou, A., Doulamis, A. et al. (2017) 'Folk dance pattern recognition over depth images acquired via Kinect sensor', *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. 33, pp.587–593.

Shi, Y. (2022) 'Dancer tracking algorithm in ethnic areas based on multifeature fusion neural network', *Wireless Communications and Mobile Computing*, Vol. 2022, pp.1–11.

Shotton, J., Sharp, T., Kipman, A. et al. (2013) 'Real-time human pose recognition in parts from single depth images', *Communications of the ACM*, Vol. 56, No. 1, pp.116–124.

Singh, S.P., Wang, L., Gupta, S. et al. (2020) 'Shallow 3D CNN for detecting acute brain hemorrhage from medical imaging sensors', *IEEE Sensors Journal*, Vol. 21, No. 13, pp.14290–14299.

Song, S., Lan, C., Xing, J. et al. (2018) 'Spatio-temporal attention-based LSTM networks for 3D action recognition and detection', *IEEE Transactions on Image Processing*, Vol. 27, No. 7, pp.3459–3471.

Wang, C., Ma, N., Ming, Y. et al. (2019) 'Classification of hyperspectral imagery with a 3D convolutional neural network and JM distance', *Advances in space research*, Vol. 64, No. 4, pp.886–899.

Wang, Y., Li, R., Wang, Z. et al. (2023) 'E3D: An efficient 3D CNN for the recognition of dairy cow's basic motion behavior', *Computers and Electronics in Agriculture*, Vol. 205, p.107607.

Xie, L. (2023) 'Rural folk dance movement recognition based on an improved MCM-SVM model in wireless sensing environment', *Journal of Sensors*, Vol. 2023, No. 1, pp.9213689.

Xing, Y. and Zhu, J. (2021) 'Deep learning-based action recognition with 3D skeleton: a survey', *CAAI Transactions on Intelligence Technology*, Vol. 6, No. 1, pp.80–92.

Xu, H., Li, L., Fang, M. et al. (2018a) 'Movement human actions recognition based on machine learning', *International Journal of Online and Biomedical Engineering (iJOE)*, Vol. 14, No. 4, pp.193–210.

Xu, Y., Cheng, J., Wang, L. et al. (2018b) 'Ensemble one-dimensional convolution neural networks for skeleton-based action recognition', *IEEE Signal Processing Letters*, Vol. 25, No. 7, pp.1044–1048.

Yang, Z., Li, Y., Yang, J. et al. (2019) 'Action recognition with spatio-temporal visual attention on skeleton image sequences', *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 29, No. 8, pp.2405–2415.

Yao, G., Lei, T. and Zhong, J. (2019) 'A review of convolutional-neural-network-based action recognition', *Pattern Recognition Letters*, Vol. 118, pp.14–22.