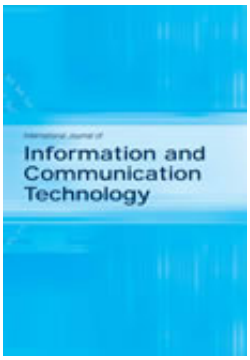
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

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

Stereo vision and deep learning-based movement correction for sports dance

Wumei Jiang, Hui Lan

Article History:

Received:	27 August 2024
Last revised:	12 September 2024
Accepted:	13 September 2024
Published online:	13 December 2024

Stereo vision and deep learning-based movement correction for sports dance

Wumei Jiang

School of Physical Education,
JiMei University,
Xiamen 361000, China
Email: 244938412@qq.com

Hui Lan*

Chengyi University College,
JiMei University,
Xiamen 361021, China
Email: lanhuijmu@126.com
*Corresponding author

Abstract: In sports dance training, dancers' wrong movements are unavoidable, and if they are not corrected in time, it will not only reduce the effect of dance expression, but also directly related to the improvement of sports performance. Therefore, we suggest a correction method for sports dance movements based on stereo vision and deep learning. Firstly, a binocular stereo imaging model is established by using the principle of triangle similarity. Secondly, 3D CNN is used to extract spatio-temporal features from the preprocessed images, and the early attention mechanism is introduced to adaptively enhance the key features that are beneficial to early action prediction. Finally, the important features are used to model the action boundaries by estimating the relative probability distribution of the action boundaries to obtain the recognition results. Simulation experiments show that the accuracy and peak signal-to-noise ratio are 91.17% and 20.45 dB, respectively.

Keywords: stereo vision; deep learning; motion correction; 3D CNN; attention mechanism.

Reference to this paper should be made as follows: Jiang, W. and Lan, H. (2024) 'Stereo vision and deep learning-based movement correction for sports dance', *Int. J. Information and Communication Technology*, Vol. 25, No. 9, pp.60–75.

Biographical notes: Wumei Jiang received her Master's degree from Guangzhou Institute of Sport in 2021. She is currently a teaching assistant of Physical Education College of Jimei University. Her research interests include dancesport, sports training methods, and cultural integration.

Hui Lan received her Master's degree from Jimei University in 2014. She is currently a Lecturer in Chengyi College of Jimei University. Her research interests include minority sports culture, dance sports education and training.

1 Introduction

As the computer vision technology rapidly growing, the correction of incorrect movements through image recognition can not only correct dancers' postures and assist dancers' training, but also be of great value to the analysis of dance techniques and accelerate the development of sport dance (Potiwetchakul, 2010). In the real physical education process, owing to the complexity and variety of sports dance movements, there is an apparent difference among the students' knowledge and understanding level and their movement capability, and some students have more wrong movements and are slower to master the correct sports dance movements (Petrenko, 2016). In this state, how to effectively rectify the incorrect movements of sport dance has become a major problem in this field. In sports dance training, incorrect movements reduce performance quality and may lead to injuries. Existing correction methods have limitations: manual correction is time-consuming and subjective, while video analysis lacks real-time feedback. The complexity of dance movements further complicates learning. Our research addresses these challenges by proposing a method for immediate, objective feedback, aiming to reduce injury risks and improve learning efficiency.

Sports dance movement correction belongs to the field of human movement recognition and classification (Potempski et al., 2022). Mallick et al. (2022) used inverse synthetic image to achieve linear characteristic point tracking and camera position approximation, combined with tracking feature points to complete the tracking and matching of linear characteristic points, so as to obtain the points with a lower degree of matching as the basis for error movement correction. Zhang et al. (2019) used different manual feature extraction for human body parts to obtain the features of dance movements, and used SVM to classify the movements.

Zhang et al. (2017) proposed a three-dimensional joint point localisation method based on stereo vision, which used Stacked-Hourglass network to detect the dancer's joint point coordinates, and localised the three-dimensional human body joint point coordinates through stereo vision system. Ji et al. (2017) used a temporal energy pyramid according to the stereo vision theory to segment the image into several small blocks, feature extraction of human target by background elimination method, obtaining three-dimensional image of human contour, solving adjacent difference frames by Laplace method, extracting key frame feature vectors, setting similarity threshold, and taking images with similarity greater than the threshold as the recognition result, but the recognition efficiency is low.

Deep learning has attracted the attention of many scholars due to its efficient performance and excellent spatial learning ability. Zhu and Zhu (2021) proposed the use of convolutional neural networks to solve the dance movement correction problem, and constructed a neural network that can learn the key features of the human body at both low and high levels, which shows the advantages of neural networks compared with the traditional two-dimensional human body gesture estimation methods. Matsuyama et al. (2021) first identify the regions where key nodes of dancers may exist, and obtain a connection graph based on all possible regions, and in this way the idea transforms the problem of correcting dance movements into a problem of classifying a dense connection graph. Yang et al. (2021) proposed feature extraction and classification of dance movement images using deep CNNs, leading to the model were poorly recognised. Lyu and Zhang (2022) designed a non-local network structure for dance action images in order to make the network model not limited to local spatial features, so that the model

distinguishes between local and non-local information and learns the global action information, but its failure to visually process the original action images resulted in poor correction.

Based on the above analysis of the current state of research, it can be seen that the existing dance movement correction methods have poor recognition accuracy due to poor image quality and insufficient key feature extraction, which affects the movement correction effect. Aiming at the above problems, this paper researches on sports dance movement correction methods through stereo vision and deep learning. The study makes significant contributions in the following key aspects.

1 Construct a binocular stereo imaging model, preprocess the acquired dance movement images, calculate the pixel values of the edge contours of the dance movement images, search for the edge of the dancer's human body contour through threshold segmentation, remove the background of the environment, lighting and other backgrounds, and eliminate the noise by using the Gaussian model, so as to provide high-quality raw images for the recognition of erroneous movements.

2 3D CNN is used for spatio-temporal feature extraction and fusion of preprocessed dance movement images to obtain stronger spatio-temporal structural representations, and the interaction of temporal features in different convolutional layers is used to enhance the characterisation of different dance movements, taking into account the need for temporal domain information.

3 The introduction of an early attention mechanism promotes the model to concentrate on the early part of the action sequence, and adaptively enhances important features for early action recognition. Important features are input into Softmax to obtain recognition results, and incorrect action correction is implemented in the form of comparisons.

4 Simulation experiments are implemented on the AIST++ dance movement dataset (Zhou et al., 2023), and the outcome indicates that the suggested method is with high recognition accuracy, peak signal-to-noise ratio, and image quality index, and can efficiently achieve the recognition and correction of dance movements.

2 Relevant technologies

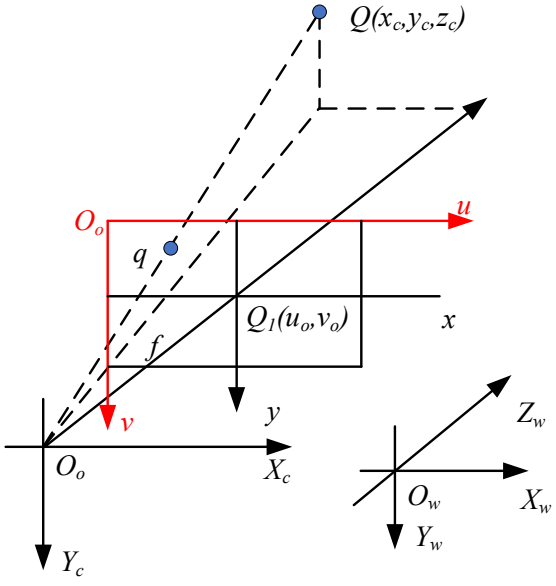
2.1 Stereo vision principle

Digital images are usually stored in the form of pixel coordinates within a medium, and the process of camera imaging is actually the process of the three-dimensional world to pixel points in the two-dimensional space of the image (Zhong and Quan, 2017). The most popularly adopted imaging model for stereo vision cameras is the pinhole model, which chiefly involves the alteration of four coordinate systems (world coordinate system, camera coordinate system, imaging plane coordinate system and pixel coordinate system) to each other (Islam et al., 2010), as indicated in Figure 1.

The planar Cartesian coordinate system is a pixel coordinate system in pixels, with origin O_0. The origin O_1 is characterised at the junction of the optical axis of the camera and the imaging plane, with coordinate (u_0, v_0). The 3D coordinate system $X_cY_cZ_c$ is the coordinate system of the camera, with the initial O_c defined in the position of the optical

centre of the camera, and $X_wY_wZ_w$ is the coordinate system of the world. O_cO_1 is the central length f of the camera, and the x-axis and y-axis of the imaging plane coordinate system $\{xy\}$ are parallel to the u-axis and v-axis in $\{uv\}$, separately. In $X_cY_cZ_c$, Z_c is the central axis of the camera, and the X_c-axis and Y_c-axis are parallel to the u-axis and v-axis of $\{uv\}$, respectively. Suppose that the world coordinates of point Q is (x_w, y_w, z_w), the coordinates of this point in the camera coordinate system is (x_c, y_c, z_c), and the coordinates of the projection of this point in the imaging plane coordinate system $\{xy\}$ is $q(x, y)$.

Figure 1 Coordinate systems in camera imaging (see online version for colours)



Supposing that the coordinates of $q(x, y)$ in the pixel coordinate system are (u, v), according to the law of similar triangles, we can get the conversion relation of point $Q(x_w, y_w, z_w)$ from 3D world coordinates to 2D pixel points as follows.

$$z_c\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f/d_x & 0 & u_0 & 0 \\ 0 & f/d_y & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}\begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} = \begin{bmatrix} M & 0_{3\times1} \end{bmatrix}\begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix}\begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \tag{1}$$

where d_x denotes the physical dimension of every pixel point in the x-axis direction of the picturing plane coordinate system, d_y denotes the physical dimension of the pixel point in the y-axis direction of the imaging plane coordinate system, and the matrices R and T are the rotation and translation transformation matrices between the world coordinate system and the camera coordinate system, respectively, and are known as the camera's outer parameter matrices. The matrix M is the intrinsic parameter matrix of the camera.

2.2 3D convolutional neural network

3D CNNs are extensions to standard convolutional neural networks (2D CNNs) that capture temporal and spatial information in images or time-series data, and therefore perform well in tasks for example image categorisation, action recognition, and prediction (Singh et al., 2019).

The size of the input level is represented by a multidimensional array of inputs (c, d, h, w), where c, d, h, and w are the amount of channels, depth, height, and width of the input data, respectively, and the size of the input level is decided in terms of the specific situation. The 3D convolutional level extracts key features from the input image. As the hierarchy of the network deepens, the convolutional level starts to extract basic features such as edges and corners from the lower level s, and then gradually transitions to the higher level s of the convolution to perform more in-depth and abstract semantic feature extraction on the basis of the previously extracted features. The output of the convolutional level is shown below:

$$Y = \frac{X + 2P - L}{S} + 1 \tag{2}$$

where X is the input, P is the padding value, L is the size of the convolution kernel, S is the step size, and Y is the convolution output.

A 3 × 3 × 3 convolution cube is divided from the 3D input, and these 27 points are convolved so that each element is multiplied by the convolution kernel and summed to obtain the output value. The input data of 3D convolutional level is difficult to represent the complete information by simple linear relationship, so the nonlinear activation functions Sigmoid, Tanh, Relu (Steinerberger and Wu, 2023) are used after the 3D convolutional level, which strengthen the 3D CNN model's ability of expressing complex data and feature learning.

Stereo vision and deep learning were chosen for their superior ability to capture spatial-temporal information and handle complex movements. Unlike motion capture or wearable sensors, this approach is non-invasive and scalable. 3D CNN excels in extracting hierarchical features from dance sequences, while stereo vision provides accurate depth perception, crucial for precise movement analysis.

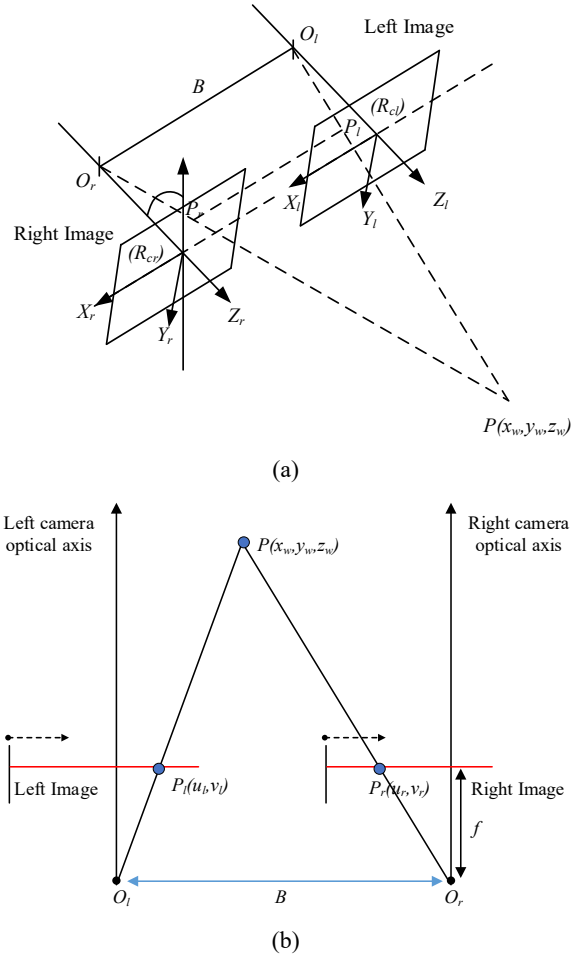
3 Binocular stereo visual imaging modelling and pre-processing of sports dance movements

3.1 Modelling binocular stereo vision imaging

In stereo vision, to gain the position of a point in space, it is essential to describe the target point in space and the camera plane, gain the association between the camera coordinates and image coordinates, and establish a binocular stereo vision imaging model. On this basis, this article do image preprocessing, calculate the pixel value of the edge contour of the dance movement image, search for the edge of the dancer's human body contour by threshold segmentation, remove the background of the environment and lighting, and eliminate the noise by adopting the Gaussian model, so as to reduce the complexity of the recognition of the dancer's movement by the visual system. The Gaussian model application in preprocessing significantly reduces background noise

while preserving edge information critical for movement detection. This step, combined with adaptive thresholding, enhances the contrast between the dancer and the background, improving the accuracy of subsequent feature extraction and movement analysis by the 3D CNN.

Figure 2 Geometric and parallax principles of binocular stereopsis, (a) the geometric principle of binocular stereo vision (b) parallax principle (see online version for colours)



From equation (1), it can be seen that the 3D coordinate (x_w, y_w, z_w) is projected as a 2D pixel coordinate (u, v) on the imaging plane, and the depth information is lost, and (u, v) actually corresponds to a straight line in 3D space, so it is not possible to recover the 3D coordinate information by using a monocular camera only (Tian et al., 2022). Therefore, stereoscopic imaging of the same target through multiple known viewpoints is an effective method to measure the 3D information of dancers.

The geometric principle of binocular stereoscopic imaging is shown in Figure 2, where B is the distance between the optical centres of the left and right cameras, and the coordinates of an articulation point P of the dancer in the coordinate system of the left camera are $P(x_l, y_l, z_l)$, and in the right camera's coordinate system are $P(x_r, y_r, z_r)$. The

forecast points of the articulation point P in the picturing planes of the left and right cameras are P_l, P_r, and the coordinates of the corresponding pixels are $P_l(u_l, v_l)$ and $P_r(u_r, v_r)$.

The X-axis of the left and right camera coordinate systems are co-linear, so there is $v_l = v_r$. From Figure 2, the pixel parallax of the hand joint point P on the left and right cameras is $d = u_l - u_r$. From the relationship established between the left and right camera coordinate systems. The equation is implied as bellow.

$$\begin{cases} x_2 = x_w - B \\ y_2 = y_w, z_w = z_2 \end{cases} \tag{3}$$

Joint equation (1) and equation (3) yield the depth information (third dimension coordinates) of the joint P as follows:

$$z_w = \frac{(x_w - x_2) f_u}{u_r - u_l} = \frac{B f_u}{d} \tag{4}$$

Substituting equation (4) into equation (1), the 3D coordinates of the joint P are calculated as follows. After obtaining the 3D coordinates, the link among the world coordinates and the image pixel coordinates of P can be obtained.

$$x_w = B\frac{u_l - u_0}{d}, y_w = B\frac{v_l - v_0}{d} \cdot \frac{f_u}{f_v}, z_w = B\frac{f}{d} \tag{5}$$

where f_u and f_v are the standardised central lengths of the camera, respectively, (u_0, v_0) is the coordinate of the principal point of the pixel coordinate system. The internal parameters of the camera are determined by the camera's camera and focal length.

3.2 Sorting target types and spatial parameters

For the goal of achieving the correction of erroneous movements in continuous sports dance, the acquired images are first pre-processed and the template combination equations for binocular stereo vision imaging of the images are calculated.

$$I(x, y) = h(x, y) * f(x, y) + \delta(x, y) \tag{6}$$

where $h(x, y)$ denotes the parallax function, $*$ describes the convolution, $f(x, y)$ is the picture edge contour pixels and $\delta(x, y)$ represents the picture sub-pixel features. Estimation of edge contour pixel values for sports dance image

$$\hat{f}(x, y) = \alpha F(x, y) + (1 - \alpha) A + \sigma^2 \tag{7}$$

where α represents the differential pixel eigenfactor, A is the pixel cloth texture set, and σ^2 represents the local area variance.

A reliable initial image is obtained by the above processing, and in order to separate the moving part of the dancer from the whole image, it is necessary to choose a proper threshold in terms of equation (8) to determine the edge of the dancer's silhouette.

$$T = T_{px,y}\left[f(x, y), p(x, y)\right] \tag{8}$$

where $f(x, y)$ represents the grey value at pixel (x, y) and $p(x, y)$ represents the grey gradient operation at this point. The segmented image of the dancer's movement can be obtained by using the above equation.

Using Gaussian mixture model (Martins et al., 2018) to remove the background, the foreground target can be obtained. For a new input image, assuming that its pixel value is Q_t, the expression for determining whether this pixel and the Gaussian model can match is as follows. If the following formula is met, the pixel and the Gaussian model match each other and belong to the background point; on the contrary, this point cannot match the model and belongs to the foreground point.

$$\left|Q_t - \mu_{i,t-1}\right| \le 2.5\delta_{i,t-1} \tag{9}$$

where $\mu_{i,t-1}$ denotes the mean and $\delta_{i,t-1}$ denotes the variance.

After Gaussian model background elimination, noise is generated near the foreground and therefore interferes with the visual system to recognise the target. So noise reduction is used to protect the image edges. The expression of the output pixel grey value after noise reduction is as bellow.

$$g(x, y) = median\left\{f(x-i, y-j)\right\}(i, j) \in W \tag{10}$$

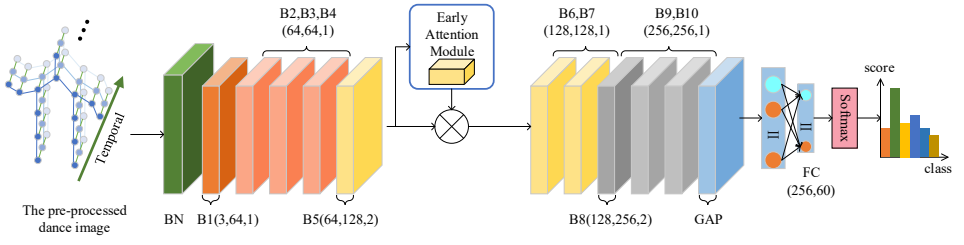
where W denotes the template window, $g(x, y)$ and $f(x - i, y - j)$ represent the output and input pixel grey values respectively.

4 Stereo vision and deep learning based movement correction for sports dance

4.1 3D CNN-based feature extraction for sports dance images

To increase the accuracy of sports dance movement recognition and improve the effect of movement correction, this paper designs a sports dance movement correction method relied on stereo vision and deep learning. Spatio-temporal features are extracted from the high-quality stereo vision images preprocessed in the previous section using 3D CNN, and the early attention mechanism is introduced to adaptively enhance the discriminative information that is beneficial for early action prediction. The critical features are inputted into Softmax to output the category prediction results of dance movements, and the comparison form is used to achieve the error movement correction. The model structure of the suggested dance correction method is shown in Figure 3.

Figure 3 The model structure of the proposed dance correction method (see online version for colours)



After obtaining the preprocessed images, this paper uses 3D CNN to extract features from the movements of sports dance images. Assume that the input vector in 3D CNN is $G \in R^{c\times t\times h\times w}$, where c is the amount of channels, h and w stand for the height and width of the feature map, respectively, and t represents the temporal dimension of the feature map. Firstly, the spatio-temporal dimension of G is done pooling operation to get the features X, $Y \in R^{c\times t}$ representing temporal and spatial information respectively. To unify the dimensions of the spatio-temporal features, the up-sampling operation of the spatio-temporal dimension is performed on X and Y respectively, and the processed spatio-temporal features are as follows.

$$X = \left[X_1, X_2, ..., X_{c_1}\right]^T, Y = \left[Y_1, Y_2, ..., Y_{c_2}\right]^T \tag{11}$$

where c_1 and c_2 represent the number of channels of spatio-temporal features, respectively. The spatio-temporal fusion is performed by a bilinear pooling operation and is calculated as follows.

$$Z_i = X^T W_i Y + b_i \tag{12}$$

where $W_i \in R^{c_1\times c_2}$ is the weight matrix of the output vector Z_i, b_i is the offset value, and Z_i is the vector of dimension i in the third vector space. Matrix decomposition of the weight matrix W_i yields the approximation matrix $W_i = U_i V_i^T$, where $U_i \in R^{c_1\times l}, V_i \in R^{c_2\times l}$. Substituting $W_i = U_i V_i^T$ into equation (12) yields the following equation.

$$Z_i = U_i^T X \cdot V_i^T Y + b_i \tag{13}$$

The fused spatio-temporal feature $Z = U^T X \cdot V^T Y + b$ is obtained by extending equation (13), where $U \in R^{c_1\times d}, V \in R^{c_2\times d}$, and $b \in R^d$ are the parameters of the bilinear pooling of the decomposition, $Z \in R^{d\times T}$ is the fused feature of X and Y, and d is the dimension of the fused new feature space.

4.2 Fine-grained feature enhancement based on attention mechanism

After obtaining the fused spatio-temporal features, the early attention mechanism (EAM) is used to generate a temporal attention map, which adaptively pays more attention to the more distinguishing message in the early content of some sequences, to improve the performance of action recognition, the structure of the EAM is implied in Figure 4. For each time step, EAM uses the time domain weights to selectively enhance features with high contribution and suppress features with less distinguishing message. The input to EAM is a d-dimensional characteristic mapping graph $Z_{in} \in R^{d\times T\times V}$, where T denotes the time step length and V denotes the amount of joints. EAM performs mean pooling of each joint to aggregate spatial message to obtain a mean pooled characteristic map $Z_{ave} \in R^{C\times T\times 1}$, which is provided to the convolutional and bulk normalisation layers to generate a set of feature maps $G_{AM}(Z_{in}) \in R^{1\times T\times 1}$.

$$G_{AM}\left(Z_{in}\right) = BN\left(W_{Conv}\left(Z_{ave}\right)\right) \tag{14}$$

where W_{Conv} denotes a one-dimensional convolution operation and BN denotes a bulk normalisation level. The feature mapping map is to be triggered by the Sigmoid level to

generate the related weights among 0 and 1. For the goal of improving the performance of movement recognition, this chapter exploits the monotonically augmenting nature of the Sigmoid activation operation to adjustably promote the model to concentrate on the early part of the action. For this purpose, a set of translation offsets is denoted by χ and initialised as [0, 1 / (T – 1), …, (T – 1) / (T – 1)]. The translation offsets are used to boost the significance of early observations. Finally, the early attention map is produced as shown in equation (15).

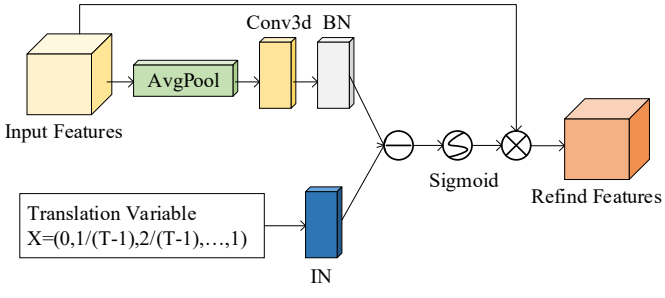
$$G_{EAM}\left(Z_{in}\right)=\delta\left(G_{AM}\left(Z_{in}\right)-IN(\chi)\right) \tag{15}$$

where δ is the Sigmoid activation operation and IN is the standardisation level adjusting the translation variables by acquirable parameters.

At last, the attention graph $G_{EAM}(Z_{in}) \in R^{1\times T\times 1}$ is subjected to an element-level multiplication operation with the input characteristic map for adjustive characteristic refinement denoted as Z_{ref}, as shown in equation (16).

$$Z_{ref}=Z_{in}*G_{EAM}\left(Z_{in}\right) \tag{16}$$

Figure 4 Early attention mechanism process (see online version for colours)



4.3 Sports dance movement recognition and correction

The fine-grained fusion feature Z_{ref} is fed into the action recognition module. Z_{ref} is processed by the start boundary branch, end boundary branch and centre offset branch in parallel, and the response intensity of each moment in the feature sequence Z_{ref} as the start boundary and the end boundary is predicted respectively, and the feature sequences Z_{start} and Z_c are obtained.

The relative boundary distribution is further estimated by summing the predicted boundary response intensity at moment t with the centre offset at the corresponding moment t on an element-by-element basis. In addition, the offset is calculated based on the current expectation value. In this case, for example, the process of predicting the response intensity of the starting boundary is expressed as equation (17) and equation (18).

$$\tilde{P}_{start}=Soft\max\left(Z_{start}^{[(t-B):t]}+Z_{c}^{t,0}\right) \tag{17}$$

$$d_{st}=E_{b\sim\tilde{P}_{start}}[b]\approx\sum_{b=0}^{B}\left(b\tilde{P}_{stb}\right) \tag{18}$$

where $\tilde{P}_{start}$ is the relative probability, which represents the probability that each moment t is the start of the action; $Z_{start}^{[(t-B):t]}$ and $Z_c^{t,0}$ represent the characteristic of moment t and the central offset predicted only by moment t, respectively; and d_{st} represents the distance from moment t to the start of the action. Similarly, the distance d_{et} from moment t to the end point of the action can be derived. The corresponding predictions of the features of each level of the 3D CNN are scaled to 2^{l-1} by the predefined local features to obtain the updated features of each level $Z^l \in R^{(2^{l-1}T)\times D}$, and the output at each moment t is denoted as $\hat{z}_t^l = (\hat{c}_t^l, \hat{d}_{st}^l, \hat{d}_{et}^l)$.

Loss function (Muhammad et al., 2021) was used in the training phase as shown bellow:

$$L = \frac{1}{N_{pos}} \sum\nolimits_{1,t} 1\{c_t^l > 0\}\left(\delta_{IOU} L_{cls} + L_{reg}\right) + \frac{1}{N_{neg}} \sum\nolimits_{1,t} 1\{c_t^l = 0\} \tag{19}$$

where δ_{IOU} is the temporal intersection ratio, L_{cls} and L_{reg} are the focal loss function and IOU loss function respectively, N_{pos} and N_{reg} are the number of positive and negative samples respectively. Based on the above results, the action boundary distances $\hat{d}_{st}^l$ and $\hat{d}_{et}^l$ are estimated to obtain the action instances a. The final predictions of the action categories as well as the start time $\hat{S}_t$ and the end time $\hat{e}_t$ are obtained in the model testing phase and the redundant instances are reduced using Soft-NMS (Chen et al., 2023) as follows:

$$a = \left(\hat{s}_t, \hat{e}_t\right) \tag{20}$$

$$\hat{s}_t = \left(t - \hat{d}_{st}^l\right) \times 2^{l-1} \tag{21}$$

$$\hat{e}_t = \left(t + \hat{d}_{et}^l\right) \times 2^{l-1} \tag{22}$$

After identifying the classes of dance movements, the error correction is implemented in the form of a comparison. A point in the input vector G is $P(x, y)$ and its affine phantom parameter on the scale σ is given by.

$$F = \begin{vmatrix} \dfrac{\partial^2 f(\sigma)}{\partial x^2} & \dfrac{\partial^2 f(\sigma)}{\partial x \partial y} \\ \dfrac{\partial^2 f(\sigma)}{\partial y \partial x} & \dfrac{\partial^2 f(\sigma)}{\partial y^2} \end{vmatrix} I_{(k)}(i, j) I(i, j) \tag{23}$$

Choosing the critical value $P(x, y)$ of the Hessian matrix, the following equation is given.

$$\frac{\partial^2 f(\sigma)}{\partial x \partial y} - \frac{\partial^2 f(\sigma)}{\partial y \partial x} = 0 \tag{24}$$

Therefore, the neighbouring frames of the current frame I_c are represented as follows.

$$NF_c = \{n; c - k \le n \le c + k\} \tag{25}$$

Calculating the average value of the error action parameters, the following error action correction factors can be obtained.

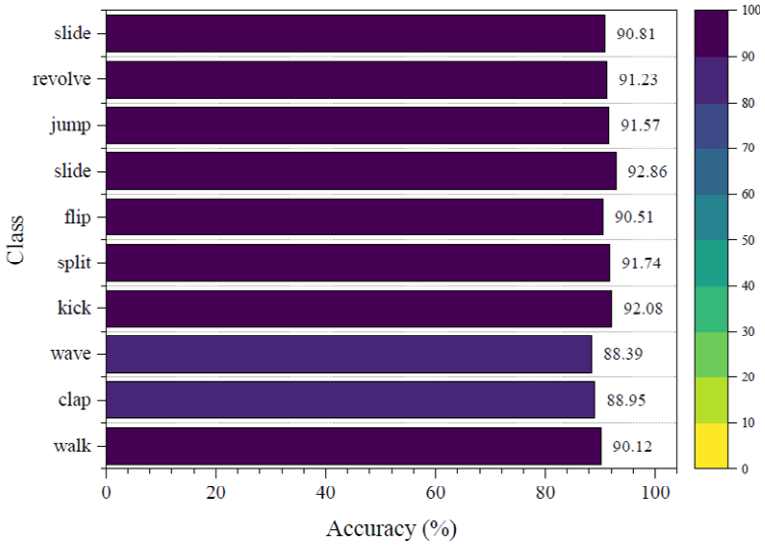
$$H = \frac{\partial^2 f(\sigma)}{\partial x^2}\frac{\partial^2 f(\sigma)}{\partial y^2} - \left(\frac{\partial^2 f(\sigma)}{\partial x \partial y}\right)^2 \tag{26}$$

5 Experimental results and analyses

To verify the recognition and correction effect of the proposed model in sports dance error movements, the AIST++ dance movement dataset (Zhou et al., 2023) was used to conduct simulation experiments, and a total of 1,579 dance images were selected, including ten common dance movements such as walking, clapping, waving, kicking, splitting, and flipping, etc., and the dataset was divided into a training set and a test set, in which the training set contained 1,105 images and the test set contains 474 images. The model is trained using Adam with an initial learning rate of 0.0001, using a cosine studying rate decay strategy with Batch size set to 2, weight decay of 0.0001, and a Soft-NMS threshold set to 0.5. The experiments are based on the Pytorch Deep Learning Platform framework, with the hardware configuration of Intel® CoreTM i5-12400 CPU, and NVIDIA 3080 GPU.

The accuracy of the proposed model for each type of movement recognition is shown in Figure 5, the recognition accuracy of walk, kick, split, flip, slide, jump, revolve and slide are all above 90%, but the recognition accuracy of clap and wave is relatively low, due to the similarity of clap and wave movements, the recognition accuracy is not as good as that of the other eight dance movements, but the overall recognition accuracy is above 90%, which verifies the effectiveness of the designed method.

Figure 5 Recognition accuracy of the proposed model for various types of dance movements (see online version for colours)



For the goal of further evaluating the recognition performance of the models, comparative experiments are conducted on SLTE (Ji et al., 2017), DCNN (Zhu and Zhu, 2021), NRDN (Yang et al., 2021), and the proposed model Ours using accuracy, TOP1 and TOP5 accuracy as the metrics, and the comparison outcome are indicated in Table 1.

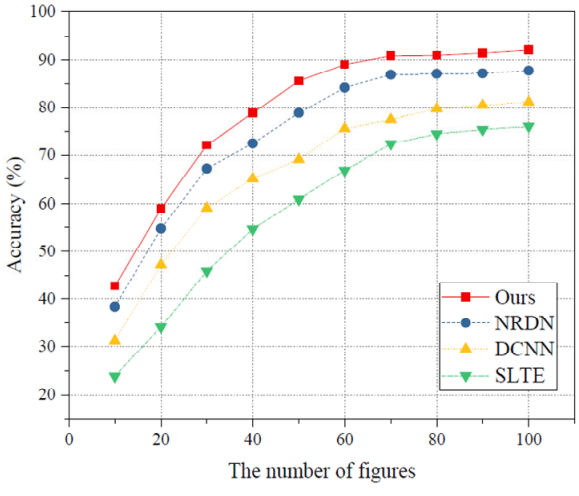
TOP1 and TOP5 accuracies are common metrics used to assess the performance of a model, and the difference lies in whether the model predicts the first few highest probability categories correctly or not. As can be seen from Table 2, the suggested model has the best accuracy, TOP1 and TOP5 accuracy metrics among all the models, which are 14.88%, 10.9% and 21.2% higher than SLTE, 9.21%, 5.2% and 10.6% higher than DCNN, and 3.53%, 2.5% and 3.7% higher than NRDN, respectively.

Table 1 Comparative test results

Model	*Accuracy (%)*	*TOP1 (%)*	*TOP5 (%)*
SLTE	76.29	40.8	50.9
DCNN	81.96	46.5	61.5
NRDN	87.64	49.2	68.4
Ours	91.17	51.7	72.1

The recognition performance of SLTE is the worst. Although the stereoscopic vision of the image is fully considered, the performance of SLTE is far lower than that of the other three models because it is only based on the traditional manual method for image feature extraction and similarity judgement as recognition results. The accuracy rate of DCNN reached more than 80%, which has certain validity. However, it only considers the transformation of dance movement images into connection graphs without stereo vision *l* and image preprocessing, resulting in insufficient feature extraction. The performance indicators of NRDN are the closest to ours, but the recognition performance of NRDN is weaker than ours due to the lack of fine-grained feature enhancement of dance movements during training. Thus, ours has obvious superiority in the recognition of sports dance movements.

Figure 6 Dance movement recognition performance comparison (see online version for colours)



The variation of recognition accuracy of different models with the number of images is shown in Figure 6. The prediction accuracies of all four models increase with the enhancement of the number of images, and ours obtains a higher performance gain because ours not only models and preprocesses the stereo vision of the dance images, but also extracts fine-grained spatial-temporal features of the dance movements using the early attention mechanism, and the learned characteristics are complementary to certain movement instances, which further indicates the high efficiency of ours for the dance movement recognition task.

In addition to the recognition performance described above, the effectiveness of the ours correction needs to be objectively examined, as measured using the peak signal-to-noise ratio (SNR), the normalised mean square error (NMSE), the multi-scale structural similarity (MS-SSIM) and the image quality index (IQI), as indicated in Table 2 for the comparisons.

Table 2 Comparison of calibration effects of different models

Model	*SNR (dB)*	*NMSE*	*MS-SSIM*	*IQI*
SLTE	9.29	0.205	0.127	0.461
DCNN	14.08	0.136	0.186	0.525
NRDN	18.26	0.081	0.228	0.568
Ours	20.45	0.067	0.243	0.581

From the above table, the SNRs of Ours, SLTE, DCNN, and NRDN are 20.45 dB, 9.29 dB, 14.08 dB, and 0.581, respectively, which indicates that the images processed by the proposed model are clearer and more conducive to recognition. MS-SSIM and IQI are proportional to the quality of stereoscopic vision, and ours has the highest MS-SSIM and IQI, with the best visual effect, making it easier to correct errors. Furthermore, the NMSE of ours is 0.067, which is reduced by 0.1983, 0.069, and 0.014 compared to SLTE, DCNN, and NRDN, respectively, indicating that the recognition error is small, and ours has high correction performance and strong generalisation ability. The recognition accuracy of clapping and waving is relatively low, which may be due to the high spatial similarity of these actions, which makes it difficult to distinguish the models. This discovery highlights the limitations of the model in dealing with subtle movement differences.

Although AIST++ dataset provides a variety of dance movements, we realise that it may not fully represent all sports dance styles and complexity. Future research will consider using multiple datasets or creating synthetic data to cover a wider range of dance styles and individual differences.

6 Conclusions

To improve the effect of sports dance movement correction, this paper researches on the basis of binocular stereo vision and deep learning. By calculating the template feature combination equation of image imaging, we obtain the pixel values of image edge contour, and then after image threshold segmentation, we can effectively distinguish the target from the background, and provide high-quality original images for dance movement recognition. Then, 3D CNN is adopted to extract spatio-temporal

characteristics from the preprocessed dance action images to obtain stronger spatio-temporal feature representations, and the early attention mechanism is introduced to encourage the model to pay more attention to the key features of the action sequences at an early stage. Finally, the critical features are input into Softmax, and the recognition results are output by estimating the relative probability distribution of the action boundaries, and the error action correction is achieved by using the contrast form.

The simulation experiments indicate that the offered method improves the recognition accuracy and efficiency, and generates important application value for dancer movement analysis. The designed method in this paper is based on binocular vision imaging for feature extraction and action recognition, and in future work, further research will be conducted on the theory of polyocular vision to improve the generalisation.

Although the experiments in this study mainly focus on specific datasets, the future work will include a wider range of actual scene tests to further verify the robustness and generalisation ability of this method. In addition, we will explore the possibility of cooperation with professional coaches and collect feedback through field application, so as to evaluate the performance of this method in actual sports dance training more comprehensively. In addition to exploring the theory of multi-vision, future research directions include developing hardware systems for real-time applications and interdisciplinary cooperation with wearable technologies. These expansions will help to transform the theoretical results of this study into practical applications and further improve the effect of sports dance training.

References

Chen, F., Zhang, L., Kang, S. et al. (2023) 'Soft-NMS-enabled YOLOv5 with SIOU for small water surface floater detection in UAV-captured images', *Sustainability*, Vol. 15, No. 14, p.10751.

Islam, A., Asikuzzaman, M., Khyam, M.O. et al. (2020) 'Stereo vision-based 3D positioning and tracking', *IEEE Access*, Vol. 8, pp.138771–138787.

Ji, X., Cheng, J., Tao, D. et al. (2017) 'The spatial Laplacian and temporal energy pyramid representation for human action recognition using depth sequences', *Knowledge-Based Systems*, Vol. 122, pp.64–74.

Lyu, Y. and Zhang, C. (2022) 'A new frog leaping algorithm-oriented fully convolutional neural network for dance motion object saliency detection', *Computer Science and Information Systems*, Vol. 19, No. 3, pp.1349–1370.

Mallick, T., Das, P.P. and Majumdar, A.K. (2022) 'Posture and sequence recognition for Bharatanatyam dance performances using machine learning approaches', *Journal of Visual Communication and Image Representation*, Vol. 87, p.103548.

Martins, I., Carvalho, P., Corte-Real, L. et al. (2018) 'Bmog: boosted Gaussian mixture model with controlled complexity for background subtraction', *Pattern Analysis and Applications*, Vol. 21, pp.641–654.

Matsuyama, H., Aoki, S., Yonezawa, T. et al. (2021) 'Deep learning for ballroom dance recognition: a temporal and trajectory-aware classification model with three-dimensional pose estimation and wearable sensing', *IEEE Sensors Journal*, Vol. 21, No. 22, pp.25437–25448.

Muhammad, K., Ullah, A., Imran, A.S. et al. (2021) 'Human action recognition using attention based LSTM network with dilated CNN features', *Future Generation Computer Systems*, Vol. 125, pp.820–830.

Petrenko, N. (2016) 'Mastering of musical rhythm by pre-school age children with speech disorders with the help of dance-correction program trainings', *Pedagogics, Psychology, Medical-biological Problems of Physical Training and Sports*, Vol. 20, No. 4, pp.23–28.

Potempski, F., Sabo, A. and Patterson, K.K. (2022) 'Quantifying music-dance synchrony during salsa dancing with a deep learning-based 2D pose estimator', *Journal of Biomechanics*, Vol. 141, p.111178.

Potiwetchakul, S. (2010) 'Correction methods of organ posture for Thai classical dancing, according to basic Thai Royal Court classical dancing standard', *Fine Arts International Journal*, Vol. 14, No. 1, pp.6–17.

Singh, R.D., Mittal, A. and Bhatia, R.K. (2019) '3D convolutional neural network for object recognition: a review', *Multimedia Tools and Applications*, Vol. 78, pp.15951–15995.

Steinerberger, S. and Wu, H-T. (2023) 'Fundamental component enhancement via adaptive nonlinear activation functions', *Applied and Computational Harmonic Analysis*, Vol. 63, pp.135–143.

Tian, X., Liu, R., Wang, Z. et al. (2022) 'High quality 3D reconstruction based on fusion of polarization imaging and binocular stereo vision', *Information Fusion*, Vol. 77, pp.19–28.

Yang, X., Lyu, Y., Sun, Y. et al. (2021) 'A new residual dense network for dance action recognition from heterogeneous view perception', *Frontiers in Neurorobotics*, Vol. 15, p.698779.

Zhang, F., Wu, T-Y., Pan, J-S. et al. (2019) 'Human motion recognition based on SVM in VR art media interaction environment', *Human-centric Computing and Information Sciences*, Vol. 9, No. 1, p.40.

Zhang, W., Liu, Z., Zhou, L. et al. (2017) 'Martial arts, dancing and sports dataset: a challenging stereo and multi-view dataset for 3d human pose estimation', *Image and Vision Computing*, Vol. 61, pp.22–39.

Zhong, F. and Quan, C. (2017) 'A single color camera stereo vision system', *IEEE Sensors Journal*, Vol. 18, No. 4, pp.1474–1482.

Zhou, Q., Li, M., Zeng, Q. et al. (2023) 'Let's all dance: enhancing amateur dance motions', *Computational Visual Media*, Vol. 9, No. 3, pp.531–550.

Zhu, F. and Zhu, R. (2021) 'Dance action recognition and pose estimation based on deep convolutional neural network', *Traitement Du Signal*, Vol. 38, No. 2, pp.1–10.