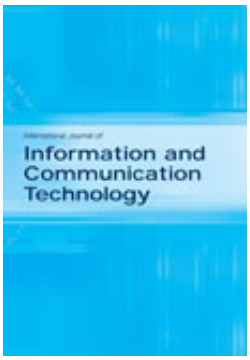


International Journal of Information and Communication Technology

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

Deep learning models combining stereo vision for dance movement evaluation

Xiaofei Ma

DOI: 10.1504/IJICT.2025.10070830

Article History:

Received:	12 March 2025
Last revised:	22 March 2025
Accepted:	22 March 2025
Published online:	06 May 2025

Deep learning models combining stereo vision for dance movement evaluation

Xiaofei Ma

School of Music,
Nanjing Xiaozhuang University,
Nanjing 211171, China
Email: mxf198456@163.com

Abstract: Computerised dance movement evaluation has grown to be a prominent research focus as computer vision and deep learning algorithms develop. Although manual annotation and 2D picture analysis are used in traditional dance movement evaluation techniques, they find it difficult to capture the dancer's 3D spatial information, therefore producing erroneous and inconsistent assessments. To address this difficulty, this work presents StereoDance-CNN-Transformer, a dance movement evaluation model leveraging stereo vision and deep learning techniques. Whereas transformer employs the self-attention mechanism for temporal modelling to capture dance movement temporal dynamics, the convolutional neural network (CNN) extracts spatial characteristics from the image and captures dance movement posture. Combining spatial and temporal data helps the model to grasp and examine difficult dancing motions. Under several dance forms, this work examined StereoDance-CNN-Transformer and showed it exceeds conventional techniques in evaluation accuracy, fluency, cross-stylistic generalisation, adaptability, and robustness.

Keywords: dance movement evaluation; stereo vision; deep learning; convolutional neural network; CNN; transformer.

Reference to this paper should be made as follows: Ma, X. (2025) 'Deep learning models combining stereo vision for dance movement evaluation', *Int. J. Information and Communication Technology*, Vol. 26, No. 11, pp.69–85.

Biographical notes: Xiaofei Ma received his Master's degree in Nanjing University of Aeronautics and Astronautics in June 2016. Currently, he works at Nanjing Xiaozhuang University. His research interests include computer vision and dance choreography.

1 Introduction

Automatic evaluation of dance movements has progressively become a research hotspot in the field of computer vision as deep learning algorithms and computer vision technologies continue to evolve (Li et al., 2024). As an art form that mostly depends on human motions, dance has tight standards on the correctness, fluidity, and timing consistency of motions. Conventional dance movement evaluation techniques mostly depend on manual annotation and video analysis, which is not only time-consuming but

also constrained by the subjective evaluation of human specialists (El Raheb et al., 2023). Research on automated dance movement evaluation has progressively attracted attention, particularly in terms of the accuracy and consistency of employing computer vision to collect and understand dance motions, after deep learning has been successfully applied in image processing and timing modelling.

Stereographic vision methods have been extensively applied in depth estimation and motion capture in 3D space recently (Clark et al., 2019). In the realm of dance movement evaluation, stereo vision can offer extensive information on the spatial gestures of dancers and assist to precisely analyse the 3D aspects of dance movements. Stereographic vision's key benefits are its ability to model human motions in 3D space after capturing the depth information of a scene from several points of view (Zhao et al., 2025). Stereoscopic vision technology is so extensively applied in multi-view image and 3D motion capture activities including motion capture in virtual reality, sports analysis, and dance movement analysis.

Although most of the conventional dance movement evaluation depends on 2D image or video data, the intricacy and dynamism of dance movements sometimes make the acquisition and analysis of 2D data insufficiently able to effectively reflect the spatial dimension. Wang et al. (2025) have tried to introduce stereo vision technology in order to address this issue. Using stereo vision technology to capture dancers' three-dimensional spatial information will yield more rich movement details than conventional two-dimensional techniques. For instance, the reconstruction of the 3D skeleton model of the human body using stereo vision can faithfully capture the spatial posture and movement trajectory of the dancer in the movement, so enhancing the accuracy and fluency of the evaluation (Nirme et al., 2020).

Apart from stereo vision, deep learning-especially transformer models and convolutional neural network (CNN) has made tremendous development in computer vision and temporal modelling. As a classic model in deep learning, CNN works well on picture classification, target identification and posture recognition (Elngar et al., 2021). It is quite successful in analysing spatial aspects in dance movements and can automatically extract low-level features in images, record information like shape, colour, and texture of objects, and. CNN does, however, have certain difficulties digesting temporal data, particularly in relation to the temporal evolution and long-term reliance of dance movements. Shaikh et al. (2024) have presented the transformer model, which has significant potential especially in action sequence modelling and multimodal data fusion and shows superiority in handling long term dependencies, therefore addressing this challenge. In the domains of action recognition, human posture estimation, and motion capture, current efforts have shown decent outcomes. For human posture estimation, for instance, CNN-based deep learning techniques may effectively extract human skeleton information from photographs; transformer has been used to temporal sequence modelling of multi-frame images, so improving the accuracy and continuity of action detection.

Still, there are several flaws in the current approaches even if certain research have made considerable progress in dance movement identification and evaluation. Though there are rather few in-depth combinations of 3D spatial analysis and time-series modelling of motions, most of the current studies centre on 2D photos. When handling complicated motions (e.g., fast-paced street dance or challenging ballet movements), most of the techniques confront the difficulties of recognition accuracy and timing consistency. Furthermore, most of the current models struggle to keep effective

performance in various dance forms while most of them lack cross-stylistic adaptability and perform just in a given dancing style.

Aiming to produce accurate assessment of dance motions using multimodal information fusion, this work presents a deep learning model, StereoDance-CNN-Transformer, which integrates stereo vision technology, CNN and transformer, to address these challenges. The StereoVision method specifically gives the model rich spatial information; the CNN is in charge of extracting the low-level picture data; and the transformer models the temporal correlations in the dance moves. By means of this creative mix, this study not only increases the generalisation capacity of the model so enabling its adaptation to the movement assessment of several dance forms, but also improves the accuracy and smoothness of dance movement evaluation.

This work has original points of interest as follows:

1 Combining stereo vision and deep learning models: the first application of merging stereo vision technology with deep learning algorithms for dance movement evaluation is shown in this work. This work is able to offer more precise spatial aspects of motions than conventional 2D photographs throughout the assessment process by using stereo vision to record 3D spatial information of dancers. This novel multi-view deep information fusion helps the model to better grasp the dancing motions in the spatial dimension, so enhancing the accuracy and fluency of the assessment.

2 Generalisation ability of multi-style dance movements: this study not only confirms the model under one dance style but also experimentally shows the outstanding generalisation capacity of the model over several dance forms. By means of cross-styles training and testing, the model is able to accommodate the movement assessment of many dance forms, so addressing the issue of significant performance variations across different dance forms and so enhancing the adaptability and flexibility of the model.

3 Innovation of integrated assessment and feedback mechanism: this work also suggests a novel integrated assessment and feedback mechanism, which incorporates evaluation indices including the fluency and consistency of the dance movements in addition to concentrating on the spatial and temporal traits of the motions. By means of this integrated assessment approach, the model can offer a more thorough and extensive study of dance moves, therefore transcending the constraints of conventional techniques emphasising just on the quality of a single movement.

2 Relevant technologies

2.1 Stereo vision technology

Stereoscopic vision technology replicating the ideas of the human visual system recovers the depth information of a 3D scene by using photos from several angles. The fundamental concept is to derive the position of the object in 3D space by means of two or more cameras obtaining pictures from various angles of the same scene and computing the parallax between them. Stereographic vision offers rich depth information, which is

appropriate for complicated dynamic scenes, and may get high-precision 3D reconstruction results by basic hardware configuration.

Image matching and triangulation provide the foundation of stereo vision's working idea. Assuming that the cameras matching two photos taken from various points of view are the left and the right cameras, first two images are acquired. First one must locate the matching spots of the same object in the left and right images to attain stereo matching. The parallax, d, is defined as the difference between the horizontal coordinates of the identical object in the left and right images: x_L, y_L for a point in the left image; x_R, y_R for a matching position in the right image.

$$d = x_L - x_R \tag{1}$$

The relative depth information of an item is reflected by parallax; the object is closer to the camera the greater the parallax value; the object is further away the smaller the parallux value. Consequently, one can immediately determine the depth information of the object by knowing the parallax's size.

From the parallax value Z, one may determine the depth of an object point by triangulation. The depth of the object point can be computed with the following formula assuming a f focal length for the camera and a B baseline distance – that is, the distance between the two cameras:

$$Z = \frac{f \cdot B}{d} \tag{2}$$

The formula reveals that the depth Z is inversely proportional to the parallax d, that is, the object is closer the larger the parallax and vice versa.

Among the main challenges in stereo vision technology is the parallax map computation. Image matching is necessary to locate the related spots in the image so obtaining an accurate parallax map (Li et al., 2022; Zhai and Chen, 2021). Usually based on pixel intensity similarity, matching techniques figure the matching cost between every pair of pixels. Common matching cost functions include more intricate correlation measurements, sum of absolute differences (SAD), and sum of squared differences (SSD). Usually, this approach presents the cost function $C(x_L, x_R)$ as the difference in intensity between the left and right pixels of the image at a given position:

$$C\left(x_L, x_R\right) = \left|I_L\left(x_L\right) - I_R\left(x_R\right)\right| \tag{3}$$

where $I_L(x_L)$ and $I_R(x_R)$ represent the pixel intensity values at respective locations x_L and x_R. This function aims to limit cost so as to get the best pixel matching relationship.

Matching techniques based on image gradient and texture characteristics have been progressively embraced recently in order to raise the accuracy of stereo matching. For instance, the computation of picture gradient can assist to identify the edge information between pixels, so improving the robustness of matching. Image gradient has a computation formula like this:

$$Grad(x, y) = \sqrt{\left(\frac{\partial I(x, y)}{\partial x}\right)^2 + \left(\frac{\partial I(x, y)}{\partial y}\right)^2} \tag{4}$$

where indicating the change in space, *Grad*(*x*, *y*) is the gradient value of the image at point (*x*, *y*). Larger gradients often suggest more obvious structural or edge changes in the image, and these areas typically have better matching accuracy.

Stereopsis vision systems additionally use parallux smoothing methods to better improve the parallax map and lower noise. Parallax smoothing aims to create a spatially continuous and smooth parallax map therefore avoiding local mistakes or noise. Common parallax smoothing techniques rely on regularising the image's gradient to prevent too significant variations in the parallax. One may formulate the goal function of this optimisation issue as:

$$L = \min_{d} \sum_{(x, y)} \left(\left(\frac{\partial d(x, y)}{\partial x} \right)^2 + \left(\frac{\partial d(x, y)}{\partial y} \right)^2 \right) \tag{5}$$

where $\left(\frac{\partial d(x, y)}{\partial x} \right)^2$ represents the parallax map's *x*-direction gradient and $\left(\frac{\partial d(x, y)}{\partial y} \right)^2$ is *y*-direction gradient accordingly. By means of the convergence of parallax values in nearby areas, the minimisation of this objective function essentially lowers noise and mistakes.

Moreover, stereo vision depends much on camera calibration (Mentzer et al., 2019). Calibrating helps one determine both internal and exterior parameters-that is, focal length, principle point position, etc., as well as relative positions between cameras. Usually utilising a known calibration plate, camera calibration is done; from the relationship between the known world coordinates and picture coordinates, one deduces both internal and external camera parameters. The following lists often used calibration formulas:

$$[x, y, 1]^T = K[R|t][X, Y, Z, 1]^T \tag{6}$$

where $[x, y, 1]^T$ are the picture coordinates; $[X, Y, Z, 1]^T$ are the coordinates of the object in the world coordinate system; *K* is the internal reference matrix of the camera; *R* and *t* are the rotation matrix and translation vector, respectively, therefore reflecting the external parameters of the camera.

Sterevision may effectively reconstruct 3D scenes and offer consistent spatial information for later movement analysis and evaluation using these methods and mathematical algorithms (Cheng and Matsuoka, 2021). Sterevision vision technology can precisely record the dynamic position changes of dancers and offer exact depth information in dance movement evaluation, therefore offering richer and more accurate data support for movement analysis.

2.2 Deep learning in dance movement assessment

Deep learning applied in dance movement assessment depends on computer vision, gesture estimation, and movement identification methods. Deep learning models-especially CNN and RNN-allow computers to automatically extract essential elements from photos or videos and conduct motion analyses to help evaluate dance quality, movement correctness, and spatial performance of the dancers. Deep learning is strong enough to properly execute movement detection and analysis in a range of

contexts and has the benefit of being able to manage challenging dynamic data in dance movement evaluation (Ferreira et al., 2021).

Common application situations in dance movement evaluation comprise tasks including movement synchronisation, pose estimation, movement classification and recognition. Deep learning – especially the processing based on time-series data – can help to support real-time evaluation of dance motions. First, a basic first stage in dance movement evaluation, posture estimation helps one examine the skeletal structure and posture alterations of the dancer. Deep learning can extract the main point information of the human body and build a 3D skeleton model by predicting the locations of several important points based on technologies such CNN and fully connected network (FCN) (Djavanshir et al., 2021).

Assume Figure 1 illustrates the breakdown of a ballet dancing movement at a specific instant.

Figure 1 Illustration of the decomposition of dance movements (see online version for colours)

Using an optimisation goal that can be characterised as the 3D location of the critical point estimated by the deep learning model:

$$\hat{p}_k = \arg\min_{p_k} \sum_i \left\| p_k - p_k^{gt} \right\|^2 \tag{7}$$

where $\hat{p}_k$ is the projected position output by the deep learning model and p_k^{gt} is the actual key point location in the calibration set. By means of this optimisation process, the deep learning model may progressively refine and precisely estimate the dancer's key point coordinates.

Second, the main activities of deep learning in dance action evaluation are action recognitions and classification. Deep learning models commonly process the time series data using RNN or LSTM for every dance movement sequence. By means of memory units, LSTM network can efficiently capture the long-term dependency of the movements in the time series, so addressing the issue of gradient disappearance that could arise for the conventional neural network handling long sequences. Assuming a dance action sequence X, one can represent X as follows:

$$X = \{x_1, x_2, ..., x_T\} \tag{8}$$

where at moment t the action feature is x. The LSTM uses this recursive algorithm to update the hidden state h_t:

$$h_t = LSTM\left(x_t, h_{t-1}\right) \tag{9}$$

where h_t, the hidden state at instant t, reflects the temporal qualities of the dance movement. The model can learn and categorise the intricate dance action characteristics using the multi-layer LSTM network. The following formula allows one to forecast the final categorisation results assuming $y \in \{1, 2, \ldots, N\}$ as the action category labels:

$$\hat{y} = \arg\max_{y} \text{soft max}\left(Wh_T + b\right) \tag{10}$$

where $\hat{y}$ is the expected action category; W and b are the model's parameters; h_T is the hidden state at the last moment; softmax(·) is the probability distribution that links the output to the category.

Recently, multimodal data fusion methods have also been included into deep learning models in order to raise the recognition accuracy (Steyaert et al., 2023). Apart from image and video data, motion capture systems or sensor data (e.g., accelerometers, gyroscopes) is frequently employed in dance movement evaluation. Deep learning models can offer a more complete movement evaluation by combining aspects of several data sources. Assuming a data input with image features f_{img} and sensor features f_{sensor}, for instance, the features can be merged via a fusion network to provide the ultimate assessment features:

$$f_{\text{final}} = \text{concat}\left(f_{\text{img}}, f_{\text{sensor}}\right) \tag{11}$$

This fused characteristic then helps to classify or evaluate actions. By means of such multimodal fusion, the restriction of a single data source may be efficiently compensated for, hence enhancing the accuracy and resilience of the model.

Apart from simple movement detection, deep learning has great application in dance movement evaluation for movement synchronisation and time alignment (Nogueira et al., 2024). Deep learning models, for instance, can synchronise movement analysis between several phases of a dance performance to guarantee that the dancers' motions match the rhythm or music. The transformer model computes the self-attention weights using the following formula assuming X as the dance movement input sequence:

$$\text{Attention}(Q, K, V) = \text{soft max}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{12}$$

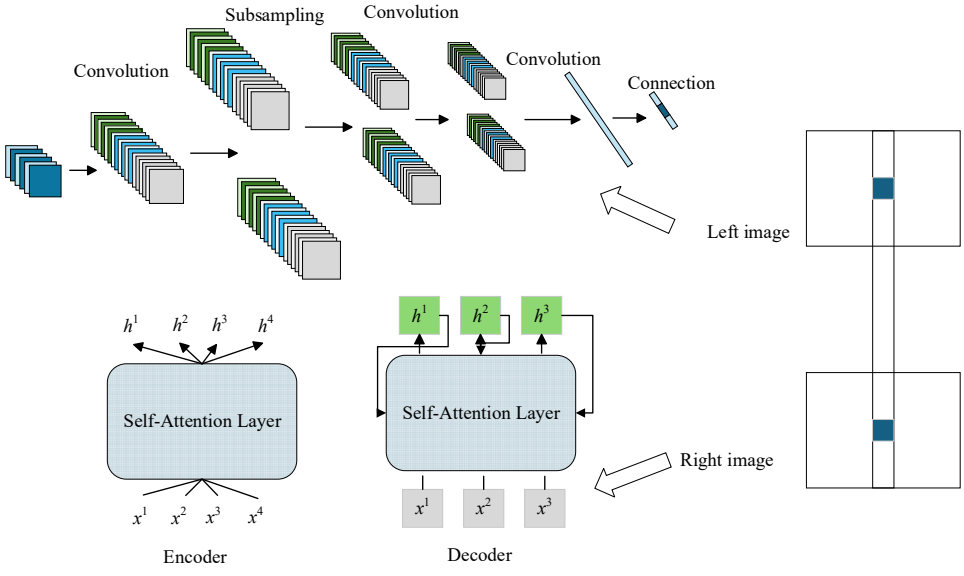
where the query vector is Q, the key vector is K; the value vector is V; the dimension of the key vector is d_k. By means of this technique, the transformer may dynamically change its attention to every time step, therefore producing more exact time synchronisation and movement analysis.

By use of several deep learning models, it can efficiently record the movement features, gesture changes, and coordination with music beat in dance, so supporting dance education, competition scoring, and real-time feedback systems.

3 StereoDance-CNN-Transformer: a stereo visual deep learning model for dance movement evaluation

Combining stereo vision techniques, CNN and transformer deep learning algorithms, the StereoDance-CNN-Transformer model seeks to precisely evaluate and analyse dance motions. Three primary components comprise the model: transformer timing modelling, StereoVision, and feature extraction and processing module. See Figure 2 to ensure the full process's complementing functions of every module guarantee the comprehensive knowledge and effective evaluation of dance moves.

Figure 2 StereoDance-CNN-Transformer model (see online version for colours)



3.1 Stereo vision module

Using several cameras, the stereo vision module records photos of dancers from various angles, therefore producing depth maps and 3D skeletal information (Pristerà et al., 2020). Recovering 3D information from 2D photos and offer correct spatial characteristics for further time-series modelling is the main goal of the module.

Stereographic matching allows one to determine the parallax between images taken from several angles. Important information regarding the depth of the scene is given by the parallax value, which reflects the variation of every pixel point in the three dimensions. One may find the depth information of every pixel by applying the parallax values. Using the 3D coordinate reconstruction formula, the stereo vision module further translates every point in the 2D image to coordinates in 3D space following the depth information. The 3D coordinate equation is:

$$X = \frac{D \cdot (x - c_x)}{f_x} \tag{13}$$

$$Y = \frac{D \cdot (y - c_y)}{f_y} \tag{14}$$

$$Z = D \tag{15}$$

The depth, D, represents the distance from each pixel point to the camera; X, Y, Z are the coordinates in 3D space; (x, y) are the pixel coordinates in the image; c_x, c_y are the photocentre positions of the camera; f_x, f_y are the focal lengths of the camera. These techniques convert each 2D pixel point to 3D coordinates, therefore generating 3D skeleton data for further action evaluation.

Processing the 3D data from consecutive frames, the stereo vision module derives the dancer's movement trajectory in space following acquisition of the 3D skeleton information. The 3D skeleton representation x_t for every frame can be obtained by first expressing these 3D skeleton points $p_{t,k}$ as the location of the dancer's joint k in 3D space at time step t:

$$x_t = \{p_{t,1}, p_{t,2}, ..., p_{t,K}\} \tag{16}$$

where K is the overall count of joints. For the next feature extraction and timing modelling modules, the 3D skeleton data supplied by the stereo vision module offers rich spatial information at last. By means of spatial location and depth information of the dancer, these data enable the model to more precisely grasp the spatial variations and structural elements of the dance motions.

3.2 Feature extraction and processing module

Extraction of meaningful spatial and temporal information from the 3D skeletal data and photos supplied by the stereo vision module is the main goal of this module; these features should be transformed into inputs fit for next temporal modelling (Yue et al., 2022). To finally produce a rich, multimodal feature representation, the module employs CNNs to extract spatial characteristics from the images and aggregates them with the depth information received from the stereo vision module.

CNNs first harvest local spatial features from the image. By layer-by- layer convolution, the CNN can extract a spectrum of representations from low-level features (e.g., edges, textures) to high-level features (e.g., forms, objects) from an image. The convolutional layers' computation is accomplished using the following equation:

$$z_i = \mathrm{ReLU}\left(\sum_j W_{ij} \cdot x_j + b_i\right) \tag{17}$$

where W_{ij} is the weight of the convolution kernel; x_j is the input feature; b_i is the bias term; z_i is the output of the convolution layer; the ReLU activation function introduces nonlinearities allowing the model to learn more intricate features.

The pooling layer then downsamples convolutional layer output to further lower feature dimensionality while maintaining significant spatial information. Usually carried out utilising maximum pooling, computed as follows:

$$p_i = \max \mathrm{pool}(z_i) \tag{18}$$

The maximum pooling process chooses the maximum value in every local area, therefore condensing the spatial dimension of the feature map and lowering the computational cost where p_i represents the pooled features.

Concurrently, this module incorporates the depth information supplied by the stereo vision module to improve the dancer's spatial awareness of her motions. The merged feature f_t is obtained by fusing the convolutional feature z_t with the depth map information d_t of every image frame:

$$f_t = \text{concat}\left(z_t, d_t\right) \tag{19}$$

The stitching procedure lets the spatial features and depth information of every image frame be fed together into the next temporal modelling module where f_t is a fused feature including depth information and spatial characteristics.

The fused features are handled through the completely connected layer to improve the expressive capability of the features even more. One may visualise this process by means of the following equation:

$$h_t = \text{ReLU}\left(W_f \cdot f_t + b_f\right) \tag{20}$$

where W_f is the weight of the completely connected layer; b_f is the bias term; h_t is the high-level feature generated by the fully connected layer. By means of the fully linked layer's processing, the dimensions of the data transfer to a new space, therefore enabling the model to learn a more abstract description of the dance motions (Kritsis et al., 2022).

This module employs CNN and stereo vision data to extract fine spatial and temporal elements in general, then feeds temporal modelling from these inputs. These characteristics not only provide rich contextual information for later dance movement evaluation but also include the 3D spatial knowledge from the depth data, so reflecting the spatial structure of the image.

3.3 Transformer timing modelling module

The transformer temporal modelling module's major goal is to use transformer deep learning architecture to replicate dance movement temporal aspects. The Module generates a high-level representation of the action by use of transformer's self-attention mechanism to capture long-range dependencies in the temporal data and absorbs processed spatio-temporal features from the Feature Extraction and Processing module. By means of this module, the model may extract efficient spatio-temporal features from intricate dance action sequences in order to accomplish correct dance action evaluation and categorisation (Qin and Meng, 2025).

First, a multi-head self-attention mechanism helps the transformer architecture understand the significant temporal aspects in the input sequence. The self-attention mechanism dynamically changes the weights of several points in the input sequence by computing their correlation. The self-attention computation method for the input feature sequence $X = [x_1, x_2, \ldots, x_T]$ follows this equation:

$$A = \text{soft}\max\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{21}$$

With Q as the query matrix, K as the key matrix, and d_k as the dimension of the key vector, Calculating the similarity between the query and the keys allows the self-attention mechanism to acquire the weights; subsequently, it uses weighted summation on the value matrix V to generate the output features for every position.

Transformer then employs a multi-head attention mechanism to process the inputs in parallel, hence enhancing the expressiveness of the model. Every head learns a different subspace representation and then stitches together the outputs of several heads to produce a richer feature representation. Multi-head attention produces outputs that might be expressed as:

$$Z = \text{concat}\left(\text{head}_1, \text{head}_2, ..., \text{head}_h\right) \cdot W_o \tag{22}$$

$$\text{head}_i = \text{Attention}\left(Q_i, K_i, V_i\right) \tag{23}$$

where head_i is the result of the i^{th} head; W_o is the output weight matrix; h is the number of heads; and the produced feature is the last result. Z combines knowledge gleaned from several subspaces learned from several brains.

Transformer's encoder also features a component for positional coding to add positional information for every input sequence position. Transformer itself lacks the capacity to manage sequence order, thus position encoding embeds positions into the input features overcomes this issue. Position encoding has a formula:

$$PE(t, 2i) = \sin\left(\frac{t}{10,000^{2i/d}}\right) \tag{24}$$

$$PE(t, 2i+1) = \cos\left(\frac{t}{10,000^{2i/d}}\right) \tag{25}$$

where i is the location coding dimension index, t is the time step; d is its dimension. Input data is location encoded so the model may manage timing data with position knowledge.

Decoder following multi-layer encoder processing will finalise processing and classification of transformer output features. The transformer timing modelling module may so extract high-level characteristics with timing patterns and capture timing dependencies in dance movement sequences to evaluate and analyse dancing stirrup movements.

This chapter describes the StereoDance-Transformer, a deep learning model with CNN, transformer, and stereo vision that precisely evaluates dancing actions. The stereo vision module first provides the model 3D spatial information needed for improved dance motion interpretation. CNN may then integrate spatial elements extracted from dancing video frames with depth information to produce rich input characteristics. Finally, transformer's temporal modelling lets it detect long-distance dependencies in dance action sequences thereby improving action evaluation and classification.

These three techniques enable the StereoDance-Transformer model to grasp spatial features and manage temporal information, therefore providing a more whole dance movement evaluation solution. Particularly in dance instruction, performance evaluation, and movement recognition, the method has many pragmatic applications.

4 Experimental results and analyses

4.1 Experimental data

The experimental dataset of this work is the multiview dance dataset. Designed for dance movement evaluation, this dataset comprises multiview video footage and matching 3D motion data, which is ideally fit for tasks integrating stereo vision and deep learning models.

Every video in the dataset is filmed from a distinct perspective to guarantee sufficient variation and stereo information to facilitate the training of stereo vision systems. Every video has several frames, and each one offers 2D and 3D posture information of the dancer that may be utilised to train models of dance movement recognition and evaluation. Table 1 compiles the dataset's key facts:

Table 1 Multiview dance dataset information

Item	*Details*
Data type	Multi-view video data, 3D human pose data
Dance styles	Includes various styles such as modern dance, ballet, and street dance
Video count	Approximately 200 videos
Frame rate	30 frames per second
Number of views	At least 4 different viewpoints per video
Annotations	2D and 3D joint coordinates for each video frame

4.2 Experimental evaluation

Several important evaluation metrics are chosen in this work to assess the StereoDance-CNN-Transformer model in dance movement evaluation, which can sufficiently evaluate the model's capabilities in terms of temporal modelling, spatial localisation and movement fluency. The particular evaluation criteria consist as follows:

4.2.1 Temporal consistency score

When running dance movement sequences, the stability and coherence of the model are assessed using the timing consistency score (Aristidou et al., 2022). Good temporal consistency helps the model to move fluidly between frames, therefore preventing sharp transitions or breaks. Calculating the similarity between the projected trajectory of the model and the actual trajectory with the formula helps one evaluate this metric:

$$TCS = \frac{1}{N}\sum_{i=1}^{N}\left|\text{Prediction}n_i - \text{GroundTruth}_i\right| \tag{26}$$

where N is the total number of video frames; Prediction_i and GroundTruth_i respectively refer to the expected and true values of the ith frame.

4.2.2 Spatial localisation accuracy

The degree of localising dancer joint points in 3D space is evaluated using accuracy (Bera et al., 2023). Analysis of dance movement depends on precise 3D joint prediction. The Euclidean distance between the expected 3D joint coordinates and the actual joint coordinates-defined as-measures this statistic:

$$SLA = \frac{1}{M}\sum_{i=1}^{M}\sqrt{\left(x_i - x_i^*\right)^2 + \left(y_i - y_i^*\right)^2 + \left(z_i - z_i^*\right)^2} \tag{27}$$

where M is the total number of joints; (x_i, y_i, z_i) and (x_i^*, y_i^*, z_i^*) correspondingly indicate the 3D coordinates of the projected and real joints.

4.2.3 Action smoothness score

The model's smoothness is assessed in producing dancing motions using action smoothness score (ASS). A key component of dance movement recognition, action smoothness tells if the change between movements is natural or not. ASS is assessed using the formula:

$$ASS = \frac{1}{N-1}\sum_{i=2}^{N}\left|\frac{v_i - v_{i-1}}{v_i}\right| \tag{28}$$

where v_i is the i^{th} frame's speed here. Reduced ASS values show more natural and smooth movements the model generates.

4.3 Experimental procedure

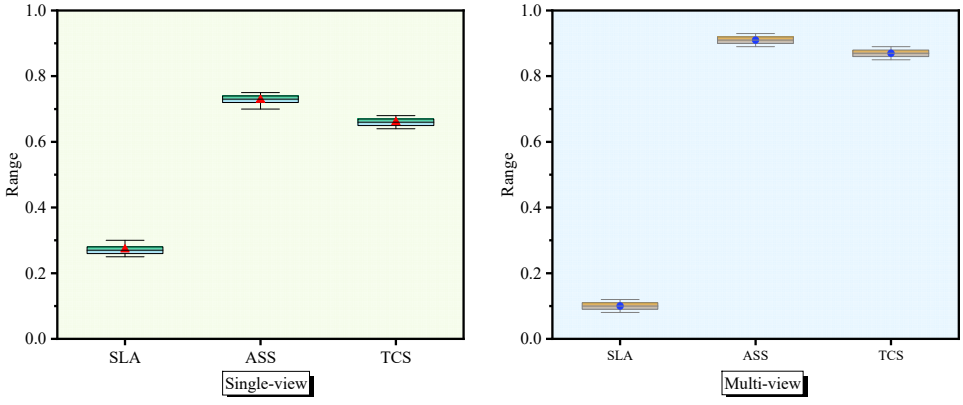
Two studies to assess the StereoDance-CNN-Transformer model for dance motion evaluation are presented in this chapter. The first experiment assesses the performance across dance forms and the accuracy of the model on a standard set of dancing actions. Modern, street, ballet, and Latin dance forms all test the model's accuracy and temporal consistency. The second experiment evaluates the generalisability and robustness of the model among dancing forms. These two experiments taken together show the capacity of the model to evaluate different dancing motions.

StereoDance-CNN-Transformer model for single-view and multi-view dance movement assessment is tested in first experiment. From every video in the collection, single-view and multi-view dancing frames are separated. Whereas the single-view condition assesses each video from a fixed viewpoint, the multi-view condition lets the model simultaneously examine video data from several camera angles.

During preparation, all of the video data was standardised. CNNs extracted keypoints and early features from every video frame for single-view data, which the transformer model for temporal modelling was fed. Sterevision vision processing was applied for multi-view data to get the 3D spatial coordinates of every dance movement by fusion calculation of several camera angles fed into the CNN for feature extraction and temporally modelled using the transformer model. Adam optimiser parameter updates and a cross-entropy loss function guides training of the model. Training makes use of a

batch gradient descent method to guarantee model generalisation across several viewpoints. The experimental results show in Figure 3.

Figure 3 Comparison results of the performance of dance movement evaluation under single view and multi-view perspectives (see online version for colours)



The model performs notably better in the multi-view condition than in the single-view condition, according to the experimental findings. The multi-view scenario produced lower scores on the spatial localisation accuracy (SLA) measurements, suggesting that the model was more spatial localised correctly. Furthermore, the much higher multi-view condition scores of ASS and temporal consistency score (TCS) show that, with multi-view, the model can better grasp the fluency and timing consistency of the dance motions.

By means of this investigation, the significant benefit of multi-view input over single-view input in dance movement evaluation was confirmed to yield more accurate and reliable assessment findings.

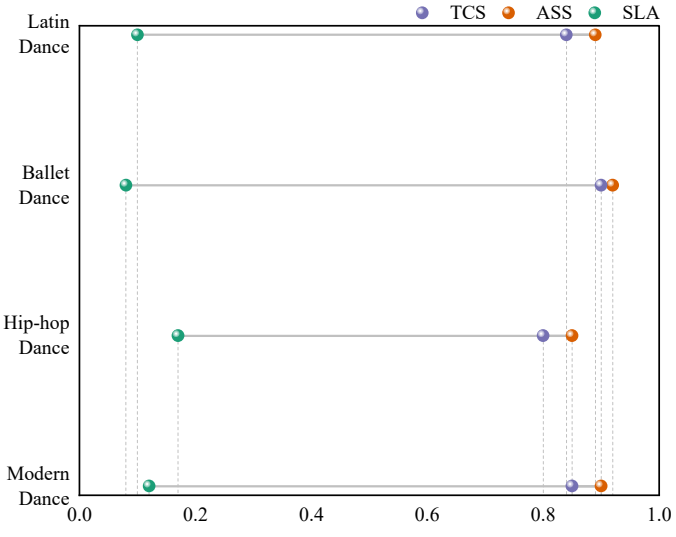
Experiment 2 assessed the StereoDance-CNN-Transformer model's generalisation capacity on the multiview dance dataset including several dance forms. This dataset consists of Multiview Dance movies spanning several dance forms; so, this experiment intends to investigate the variations in the performance of the model under several dance forms and evaluate its adaptability among several dance styles.

All videos are first normalised and then arranged based on dance forms during data preprocessing. CNN helps to extract important aspects of every video frame for every batch of data. The retrieved features are then fed into the transformer model for temporal modelling, and lastly the model produces assessment results for the dancing motions. Figure 4 shows the experimental findings.

Modern dance's experimental SLA score of 0.12 shows better model placement. With high ASS and TCS of 0.90 and 0.85, modern dance indicated strong movement fluency, timing consistency, and smoother, more steady transitions. Modern dancing fits the approach really nicely.

Street dance has a high SLA score of 0.17, but the complexity of the movements and the significant rhythmic changes lower fluency and temporal consistency and make it difficult for the model to capture smooth transitions and temporal coherence. This reflects in the ASS and TCS scores of just 0.85 and 0.80.

Figure 4 Comparison of the model's evaluation performance in different dance styles (see online version for colours)



With a lowest SLA score of 0.08, ballet shows good placement accuracy. Ballet is steady and consistent; the ASS and TCS are 0.92 and 0.90, showing smooth movements and great temporal consistency.

With an ASS score of 0.89, showing great fluency, Latin dance has a somewhat higher SLA of 0.10 than ballet; yet, a TCS score of 0.84 is probably the result of the significant rhythmic variations in the motions, which provide worse timing consistency.

5 Conclusions

StereoDance-CNN-Transformer is a dance movement evaluation model based on stereo vision and deep learning techniques (CNN and transformer). The model combines CNN's strong feature extraction with transformer's temporal modelling to precisely analyse dancing actions from several angles using stereo vision to record their spatial information.

Though it has certain results, StereoDance-CNN-Transformer has some dance movement evaluation limits. First, in complex dance genres including street dance, the model's temporal consistency and movement fluidity are lacking. Second, this work used the multiview dance dataset for experiments; although its scope and variety might restrict model evaluation, this is still a valid approach. The model problems with visual quality and motion blur even with stereo vision.

The following areas call more investigation:

1 Expanding the dataset and diversity enhancement: the dataset could include more dancing forms and movement techniques in next research. More dance footage in diverse settings and lighting conditions will help the model to be more robust and generalising. Furthermore enhancing the cross-cultural adaption of the concept is the

cultural variety of dancing motions, integrating dance forms from several civilisations.

2 Optimisation of stereo vision algorithms: future studies can investigate more sophisticated stereo vision algorithms, like stereo matching techniques based on deep learning, or the application of multi-view image fusion technologies to raise the depth map's accuracy. Stability and robustness of the model in complicated dynamic situations can be raised by optimising stereo vision algorithms.

3 Multimodal learning and self-supervised learning: future studies can incorporate multimodal learning to integrate several kinds of data, including audio, sensor data, and motion capture data, thereby offering a more complete dance evaluation model. Furthermore, a self-supervised learning method – using unlabelled data for training – can considerably raise the model's performance in an unsupervised setting and increase its adaptability.

Finally, this work shows the possibilities of merging stereo vision technology with deep learning algorithms in this field and offers a fresh perspective for evaluation of dance movements. As technology develops constantly, it is projected to enhance the performance of the model and increase its application range in the next years.

Declarations

All authors declare that they have no conflicts of interest.

References

Aristidou, A., Yiannakidis, A., Aberman, K., Cohen-Or, D., Shamir, A. and Chrysanthou, Y. (2022) 'Rhythm is a dancer: Music-driven motion synthesis with global structure', *IEEE Transactions on Visualization and Computer Graphics*, Vol. 29, No. 8, pp.3519–3534.

Bera, A., Nasipuri, M., Krejcar, O. and Bhattacharjee, D. (2023) 'Fine-grained sports, yoga, and dance postures recognition: a benchmark analysis', *IEEE Transactions on Instrumentation and Measurement*, Vol. 72, pp.1–13.

Cheng, M-L. and Matsuoka, M. (2021) 'Extracting three-dimensional (3D) spatial information from sequential oblique unmanned aerial system (UAS) imagery for digital surface modeling', *International Journal of Remote Sensing*, Vol. 42, No. 5, pp.1643–1663.

Clark, R.A., Mentiplay, B.F., Hough, E. and Pua, Y.H. (2019) 'Three-dimensional cameras and skeleton pose tracking for physical function assessment: a review of uses, validity, current developments and Kinect alternatives', *Gait & Posture*, Vol. 68, pp.193–200.

Djavanshir, G.R., Chen, X. and Yang, W. (2021) 'A review of artificial intelligence's neural networks (deep learning) applications in medical diagnosis and prediction', *IT Professional*, Vol. 23, No. 3, pp.58–62.

El Raheb, K., Buccoli, M., Zanoni, M., Katifori, A., Kasomoulis, A., Sarti, A. and Ioannidis, Y. (2023) 'Towards a general framework for the annotation of dance motion sequences: a framework and toolkit for collecting movement descriptions as ground-truth datasets', *Multimedia Tools and Applications*, Vol. 82, No. 3, pp.3363–3395.

Elngar, A.A., Arafa, M., Fathy, A., Moustafa, B., Mahmoud, O., Shaban, M. and Fawzy, N. (2021) 'Image classification based on CNN: a survey', *Journal of Cybersecurity and Information Management*, Vol. 6, No. 1, pp.18–50.

Ferreira, J.P., Coutinho, T.M., Gomes, T.L., Neto, J.F., Azevedo, R., Martins, R. and Nascimento, E.R. (2021) 'Learning to dance: a graph convolutional adversarial network to generate realistic dance motions from audio', *Computers & Graphics*, Vol. 94, pp.11–21.

Kritsis, K., Gkiokas, A., Pikrakis, A. and Katsouros, V. (2022) 'Danceconv: dance motion generation with convolutional networks', *IEEE Access*, Vol. 10, pp.44982–45000.

Li, C., Yun, L. and Xu, S. (2022) 'Blind stereoscopic image quality assessment using 3D saliency selected binocular perception and 3D convolutional neural network', *Multimedia Tools and Applications*, Vol. 81, No. 13, pp.18437–18455.

Li, J., Miao, Q., Zou, Z., Gao, H., Zhang, L., Li, Z. and Wang, N. (2024) 'A review of computer vision-based monitoring approaches for construction workers' work-related behaviors', *IEEE Access*, Vol. 12, pp.7134–7155.

Mentzer, N., Mahr, J., Paya-Vaya, G. and Blume, H. (2019) 'Online stereo camera calibration for automotive vision based on HW-accelerated A-KAZE-feature extraction', *Journal of Systems Architecture*, Vol. 97, pp.335–348.

Nirme, J., Haake, M., Gulz, A. and Gullberg, M. (2020) 'Motion capture-based animated characters for the study of speech-gesture integration', *Behavior Research Methods*, Vol. 52, pp.1339–1354.

Nogueira, M.R., Menezes, P. and Maçãs de Carvalho, J. (2024) 'Exploring the impact of machine learning on dance performance: a systematic review', *International Journal of Performance Arts and Digital Media*, Vol. 20, No. 1, pp.60–109.

Pristerà, F., Gallo, A., Fregola, S. and Merola, A. (2020) 'Development of a biomechatronic device for motion analysis through a rgb-d camera', *Global Clinical Engineering Journal*, Vol. 2, No. 3, pp.35–44.

Qin, W. and Meng, J. (2025) 'The research on dance motion quality evaluation based on spatiotemporal convolutional neural networks', *Alexandria Engineering Journal*, Vol. 114, pp.46–54.

Shaikh, M.B., Chai, D., Islam, S.M.S. and Akhtar, N. (2024) 'From CNNs to transformers in multimodal human action recognition: a survey', *ACM Transactions on Multimedia Computing, Communications and Applications*, Vol. 20, No. 8, pp.1–24.

Steyaert, S., Pizurica, M., Nagaraj, D., Khandelwal, P., Hernandez-Boussard, T., Gentles, A.J. and Gevaert, O. (2023) 'Multimodal data fusion for cancer biomarker discovery with deep learning', *Nature Machine Intelligence*, Vol. 5, No. 4, pp.351–362.

Wang, Y., Bi, W., Liu, X. and Wang, Y. (2025) 'Overcoming single-technology limitations in digital heritage preservation: a study of the LiPhoScan 3D reconstruction model', *Alexandria Engineering Journal*, Vol. 119, pp.518–530.

Yue, R., Tian, Z. and Du, S. (2022) 'Action recognition based on RGB and skeleton data sets: a survey', *Neurocomputing*, Vol. 512, pp.287–306.

Zhai, L. and Chen, D. (2021) 'Image real-time augmented reality technology based on spatial color and depth consistency', *Journal of Real-Time Image Processing*, Vol. 18, No. 2, pp.369–377.

Zhao, Y., Qin, H., Xu, L., Yu, H. and Chen, Y. (2025) 'A review of deep learning-based stereo vision techniques for phenotype feature and behavioral analysis of fish in aquaculture', *Artificial Intelligence Review*, Vol. 58, No. 1, pp.1–61.