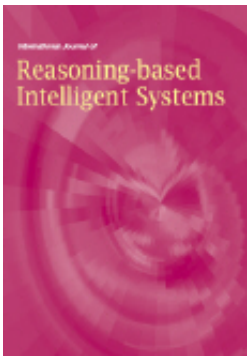


International Journal of Reasoning-based Intelligent Systems

ISSN online: 1755-0564 - ISSN print: 1755-0556
https://www.inderscience.com/ijris

Adaptive convolutional network and transfer learning-based dance movement recognition for realistic subjects

Lingmei Hu

DOI: 10.1504/IJRIS.2025.10072538

Article History:

Received:	06 May 2025
Last revised:	24 May 2025
Accepted:	24 May 2025
Published online:	24 July 2025

Adaptive convolutional network and transfer learning-based dance movement recognition for realistic subjects

Lingmei Hu

School of Music and Dance,
Hunan University of Science and Engineering,
Yongzhou 425199, China
Email: hlm2025428@163.com

Abstract: As deep learning technology develops rapidly, dance action recognition, a crucial computer vision research direction, has been progressively applied in other disciplines. But classical action recognition techniques often fail in complicated contexts with various action categories and dynamically shifting backgrounds; dance action recognition for realistic topics faces many difficulties. This work thus suggests a dance action recognition model ACN-TL-DAR based on adaptive convolutional network and transfer learning (TL), which combines adaptive convolutional networks and TL to efficiently manage complicated dance action data. This work confirms the great performance of the ACN-TL-DAR model on several criteria by means of experimental evaluation on two datasets. The experimental results reveal that the model suggested in this work has strong robustness and efficient identification capacity in several contexts, thereby offering a fresh concept for the expansion of the field of realistic dance movement recognition.

Keywords: realistic dance movement recognition; adaptive convolutional network; transfer learning; TL; temporal consistency; category balance.

Reference to this paper should be made as follows: Hu, L. (2025) 'Adaptive convolutional network and transfer learning-based dance movement recognition for realistic subjects', *Int. J. Reasoning-based Intelligent Systems*, Vol. 17, No. 9, pp.23–33.

Biographical notes: Lingmei Hu received her master degree from North Chiang Mai University in June 2023. She is currently working at Hunan University of Science and Engineering. Her research interests include machine learning, folk dance and movement recognition.

1 Introduction

Action recognition technology has been extensively applied in the domains of security monitoring, sports analysis, and intelligent interaction with the fast expansion of artificial intelligence and computer vision technologies. Usually, action recognition activities demand the system to precisely analyse and comprehend human action postures and their development (Kong and Fu, 2022). But in the particular choreography of dance movement identification, in addition to identifying fundamental movement postures and timing correlations, complicated elements like personal variations, diversity of dance forms, and backdrop environment changes have to be considered (Newell, 2020). Thus, dance movement recognition is not only a technical difficulty but also a wonderful test of the capacity of the algorithmic model to be used in the actual world.

Realistic dance movement identification confronts more difficult difficulties than conventional dance movement identification. Realistic dances can have more varied dance forms, more challenging movement combinations, and more background changes; so, the system must have great generalising capacity and resilience. Shallow machine learning techniques such support vector machine (SVM), decision tree and random forest (RF), and manual feature extraction define traditional dance movement detection approaches. To get better classification results, SVM as an example must manually choose features and transfer them to a high-dimensional space (Borji et al., 2023). When faced with complex dance movements, this approach is susceptible to noise and background interference, which diminishes the accuracy of feature extraction and subsequently affects the model's performance. Although RF and other integration techniques can enhance classification accuracy, they still rely on manually designed features and are prone to overfitting due to significant computational overhead when dealing with large-scale data. Furthermore, these conventional techniques typically cannot adequately depict the spatio-temporal link between actions, which leads to suboptimal performance in complicated situations or continuous actions.

Deep learning models' emergence in recent years has brought to notable advancement in dance movement identification. Thanks to their strong automatic feature learning capacity, convolutional neural networks (CNNs) have become among the most often utilised deep learning

architectures (Alzubaidi et al., 2021). Dance motions with temporal aspects are processed using recurrent neural networks (RNNs) including gated recurrent units (GRUs) and long short-term memory networks (LSTMs). Traditional deep learning models find it challenging to reach the intended outcomes in real-life subject matter, especially in the face of dynamic changes, complicated backgrounds and significant individual variances, even if these approaches work well in stationary or simplified settings (Panadeiro et al., 2021).

Researchers have started investigating more adaptable and flexible deep learning techniques to help to address these challenges. As a developing technology, adaptive convolutional networks introduce a dynamic tuning mechanism to dynamically change the shape and size of the convolutional kernel depending on various input data, so enabling more precisely to capture spatio-temporal aspects in complicated operations (Li et al., 2022). Furthermore, extensively applied in computer vision activities is TL as a useful learning tool. In dance action recognition, TL solves the issue of inadequate data annotation in realistic subject dances, moves information from current large-scale relevant datasets, and increases the generalisation capacity of the model in new settings. The model can be trained on limited labelled data with TL and achieve greater performance in actual applications.

With an aim to enhance the performance of the model in dance movements of realistic subjects, this work presents a dance movement recognition model based on adaptive convolutional networks and TL. More especially, this model's inventiveness shows in the following features:

1 The introduction of adaptive convolutional network: this work presents an adaptive convolutional network, which can dynamically change the size and form of the convolutional kernel to better fit the diversity and variability of complicated movements, hence enhancing the accuracy and resilience of movement detection.

2 Effective application of TL: this work uses the TL technique to migrate current information from large-scale datasets, therefore improving the recognition capacity of the model in small-sample tasks, particularly in the application of the particular domain of dance motions.

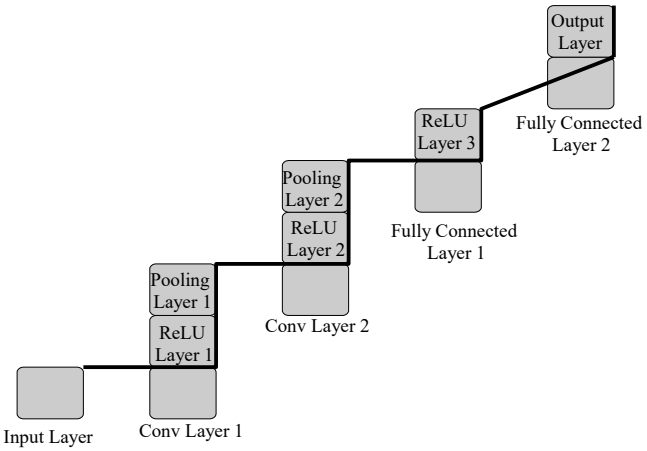
3 Multi-dimensional metrics for comprehensively evaluating model performance: three important metrics are proposed in this work in the trials to assess the performance of the ACN-TL-DAR model from several angles. This creative assessment approach not only shows the whole performance of the model but also offers a clear road for next development and optimisation.

4 In-depth analysis of ablation experiments: this work methodically investigates the particular contributions of the adaptive convolutional network and the TL module to the model performance via ablation experiments. This thorough investigation shows the synergy among the modules, which clarifies the fundamental mechanism of the model and offers a foundation for next model optimisation.

2 Adaptive convolution network

By use of a convolutional kernel adaptation mechanism to precisely capture spatio-temporal information in images, adaptive convolutional networks are able to dynamically change convolutional operations over various input data. Conventional CNNs employ fixed size and shape convolutional kernels, which causes the model to often find it difficult to effectively extract features given complex and varied inputs (Liu et al., 2021a). Multiple convolutional layers, a ReLU activation layer, a pooling layer, a fully connected layer, and an output layer define a typical CNN architecture shown in Figure 1.

Figure 1 Typical CNN architecture



CNN generates features by means of several convolutional layers and ReLU activation layers, as depicted in Figure 1; typically, a ReLU layer follows each of these layers to boost nonlinearities. Reducing the spatial dimensionality of the feature map by means of the pooling layer helps to improve the generalisation of the model by lowering the computation required. Extracted features are mapped to the output space using the fully connected layer; at last, the output layer generates the final prediction. One can characterise the convolution procedure of a standard CNN as follows:

$$Y = X * K + b \tag{1}$$

where X is the input data; K is a set convolutional kernel; b is a bias term; Y is the output feature map. The CNN cannot dynamically change the convolution kernel depending on different inputs since the convolution kernel shares parameters throughout the input data, so less accurate extraction of complex features may result.

Adaptive convolutional networks dynamically create convolutional kernels by include a generator network, therefore solving this problem (Zhang et al., 2021). Specifically, the adaptive convolutional network dynamically creates the convolutional kernel K_{gen} using a

generator network driven on the input data X. The convolution operation can so be stated as:

$$Y = X * K_{gen} + b \tag{2}$$

Under this architecture, the convolutional kernel K_{gen} is influenced by the features of the input data. Consequently, each time different data is input, the convolutional kernel can alter its form and parameters based on the data, thereby enhancing flexibility and accuracy in feature extraction.

By weighting various areas of the input data, researchers have developed an attention mechanism to improve the focus of adaptive convolutional networks on significant features, hence augmenting their performance (Li et al., 2020). In particular, the attention weight matrix A weights various points of the input data such that the network can concentrate on those areas most critical to the final output. One can expand the convolution operation and represent it as:

$$Y = \sum_{i=1}^{H} \sum_{j=1}^{W} A_{i,j} \cdot X[i, j] * K_{gen} + b \tag{3}$$

where $A_{i,j}$ is the attentional weight of every point in the input data; H and W are the height and width of the input image accordingly. By means of this weighting method, the network enhances the processing capacity for challenging jobs and better extracts features with great correlation.

The adaptive convolutional network maximises the creation of convolutional kernels during the training process by use of a multi-task learning paradigm. The total loss function can be stated assuming that the loss of the auxiliary task is L_{aux} and the loss of the main work is L_{main} as:

$$L_{total} = L_{main} + \lambda L_{aux} \tag{4}$$

where the weights of the auxiliary and main task losses are balanced using a hyperparameter λ. Through the main job, the network is able to train efficient convolutional kernels by optimising this loss function; moreover, it can offer beneficial direction through the auxiliary tasks to enhance the general performance.

Furthermore, often included to avoid overfitting and enhance the generalisation capacity of the network is a regularisation term. Assuming L_{reg} as the regularisation term, the last loss function is:

$$L_{final} = L_{total} + \gamma L_{reg} \tag{5}$$

where the regularity weight coefficient is γ. By use of regularisation, the network learns the convolutional kernel while preserving appropriate generalisation performance and thereby preventing overfitting of the training data (Liu et al., 2021b).

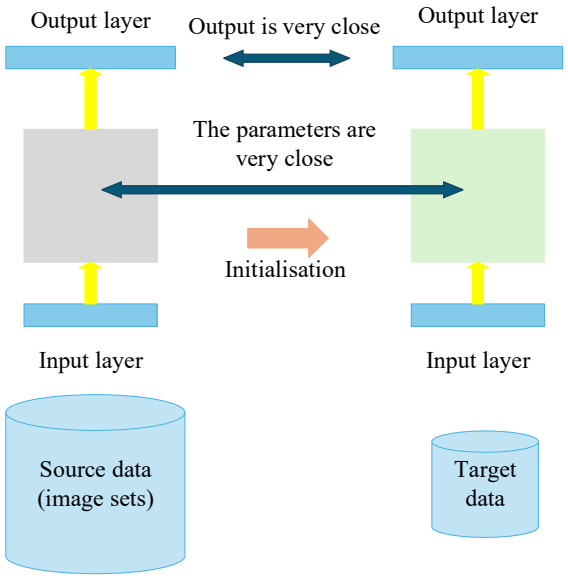
By use of the foregoing process, the adaptive convolutional network dynamically modulates the convolutional kernel on various input data, thereby enhancing the accuracy of feature extraction and displaying more flexibility in challenging tasks.

3 Transfer learning

By using the knowledge acquired in source tasks, TL is a technique to enable target task learn more efficiently. Conventional machine learning approaches hold that the target task and the source task are the same and that the training and test data originate from the same distribution (Dargan et al., 2020). In actual applications, however, the target task is often different from the source task and the target task data is usually smaller than TL might help to enhance the learning effect of the target task.

See Figure 2 for the fundamental TL framework: the models or information gained from the source task is transferred to the target task.

Figure 2 The TL framework (see online version for colours)



The source and target tasks can be shown as respectively if the dataset of the source task is D_{source} and the dataset of the target task is D_{target}:

$$D_{source} = \{X_{source}, Y_{source}\} \tag{6}$$

$$D_{target} = \{X_{target}, Y_{target}\} \tag{7}$$

The objective of the source and target tasks is to attain the best prediction by learning a model M_{target} suited to the target task where X marks the input data and Y marks the labels. Usually, TL finds the target task model M_{target} by means of fine-tuning the source task model M_{source}; this method can be shown as:

$$M_{target} = \arg\min_{\theta} L\left(X_{target}, Y_{target}; \theta\right) \tag{8}$$

where L is the loss function of the target task; θ is a target task model parameter. Pre-training the model on the source task followed by fine-tuning it on the target task helps to attain higher performance on the target task.

Effective migration relies on accurately measuring the knowledge difference between target and source tasks. A commonly used method to assess the distributional difference between source and target activities is maximum

mean discrepancy (MMD). MMD calculates the distributional difference between the target task data and the source task data using the following formula:

$$MMD^2\left(D_{source}, D_{target}\right) = \left\| \frac{1}{n}\sum_{i=1}^{n}\phi\left(X_{source}^{(i)}\right) - \frac{1}{m}\sum_{j=1}^{m}\phi\left(X_{target}^{(j)}\right) \right\|^2 \tag{9}$$

where n and m are the number of samples in the source and target task datasets respectively and $\phi(\cdot)$ is a feature mapping function transferring the input data to a high-dimensional space. Effective migration depends on reducing the MMD such that the distribution of the source and target activities differs less (Wang et al., 2022).

The usually employed method in the TL process is fine-tuning. Fine-tuning is to minimise the loss function of the target task model based on the source task model hence optimising the parameters of the target task model (Ding et al., 2023). The aim of the fine-tuning procedure can be stated as assuming the loss function of the target task is L_{target} and the model parameter of the target task is θ:

$$\theta^* = \arg\min_{\theta} L_{target}\left(X_{target}, Y_{target}; \theta\right) \tag{10}$$

By means of fine-tuning, the network can be maximised depending on the data of the target task, therefore enhancing the performance of the target task.

Using the shared feature learning method allows TL to also improve migration. Assuming the shared feature network is f_{shared}, the overall loss of the source and target jobs can be stated as:

$$L_{total} = L_{source} + L_{target} \tag{11}$$

where L_{source} and L_{target} are respectively the loss functions of the target and source jobs. By reducing the losses of both activities concurrently, shared feature learning accelerates the learning of the target task and enhances its performance (Chen et al., 2021).

Sometimes TL must change the weights of the target and source jobs to more balance the contributions of each. Assuming α for the weight of the source task and β for the weight of the target task, the overall loss may be stated as:

$$L_{final} = \alpha L_{source} + \beta L_{target} \tag{12}$$

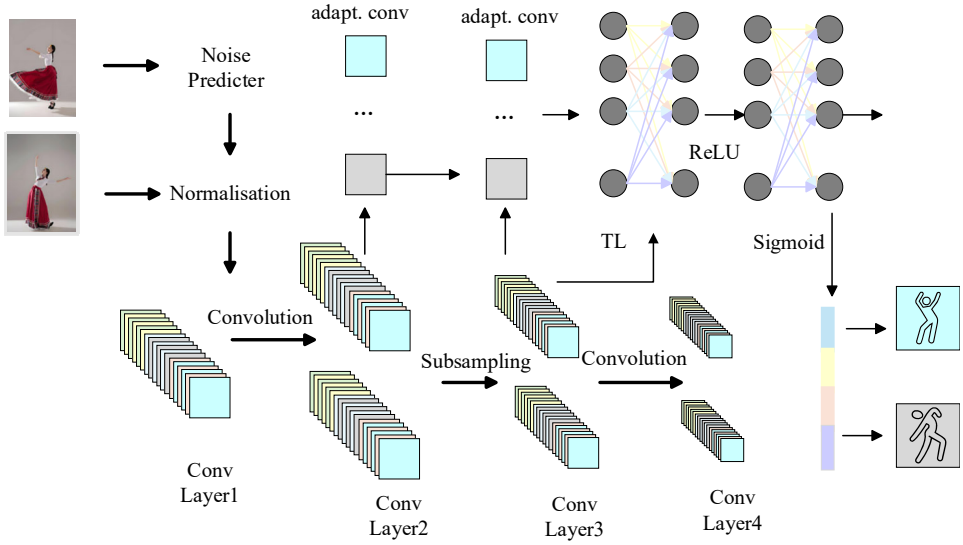
Changing the values of α and β helps to balance the effects of the source and target activities thereby maximising the migration process.

TL's basic concept is to borrow information from the source task so improving the learning effect of the target activity. Especially in cases of insufficient data for the target task, TL offers great benefits by pre-training on the source task, reducing the distributional disparities between tasks, fine-tuning the model, and sharing features, therefore improving the performance of the target job.

4 Dance movement recognition model for realistic themes

In this study, we propose a realistic dance action recognition model based on adaptive convolutional network and TL, i.e., ACN-TL-DAR. Aiming to increase the adaptability and accuracy of complicated scenarios in the task of dance action identification, the model integrates adaptive convolutional networks with TL approaches, see Figure 3. The following will go into great length on the design and execution of every module.

Figure 3 Structure of the ACN-TL-DAR model (see online version for colours)



4.1 Data preprocessing module

Data quality and processing are critical for model performance in the dance movement detection problem, particularly in real-life situations when the data sometimes includes noise, background interference, and various camera angles. Thus, the module of data preparation must guarantee that the data is consistent and clear and increase the variety to strengthen the model.

By means of picture denoising and normalisation processes, the module of data preparation enhances the data quality. Gaussian blur is one of the denosing methods applied in video frame data (Channoufi et al., 2018). Denoised image I' can be derived via convolution operation with a Gaussian kernel K assuming the input picture is I. Using a Gaussian fuzzy filter helps one to denoise the image:

$$I' = I * K \tag{13}$$

where * is the convolution operation; K is a predetermined Gaussian kernel meant to smooth the image. Denoising helps the model to significantly lower the noise interference on next feature extraction.

Second, the model also has to normalise all photos or video frames to unite the scale of the input data. Normalisation of photographs compresses the pixel values into the range [0, 1], therefore preventing learning discrepancies resulting from variations in brightness or contrast between images (Andersson et al., 2020). The normalised pixel value $p'(x, y)$ can be stated assuming that the pixel value of a picture is $p(x, y)$.

$$p'(x, y) = \frac{p(x, y) - \mu}{\sigma} \tag{14}$$

where $p(x, y)$ is the value of pixel point (x, y) in the image, μ and σ are the mean and standard deviation of the image accordingly. The model can consistently manage image data with varying illumination and contrast during the training phase by means of this normalisation operation.

4.2 Feature extraction module

The generating procedure of the convolutional kernel in an adaptive convolutional network is dynamic instead of pre-fixed. To better accommodate various dance movements, the network learns the characteristics of the input data and automatically adjusts the structure of the convolutional kernel. The generating process of the adaptive convolution kernel W can be stated assuming x as the input data:

$$W = F(x) \tag{15}$$

where $F(x)$ constructs a convolutional kernel adaptatively depending on the input data x. Learning the properties of the input data helps this function dynamically create the convolution kernel so that it may be freely changed depending on the dancing movements (Ferreira et al., 2021). The adaptive convolutional network convolves the input data with the adaptive convolution kernel across the convolution operation to produce the feature map y:

$$y = \sum_{i=1}^{N} W_i * x_i \tag{16}$$

where W_i is the i^{th} adaptive convolution kernel; $*$ is the convolution operation; x_i is the i^{th} local region of the input data; y is the extracted feature map; N is the total number of local regions of the input data.

Using a multilayer convolutional structure which lets the model extract features from low to high level layer by layer, the adaptive convolutional network also helps to improve the feature extracting capacity. Every layer's convolution kernel is dynamically changed in line with the output of the one before it. Assuming y_l as the l^{th} layer's output, one may define the convolution operation of this layer as:

$$y_l = \sum_{j=1}^{M} W_l^j * y_{l-1}^j \tag{17}$$

where W_l^j is the j^{th} adaptive convolution kernel of the l^{th} layer; y_{l-1}^j is the j^{th} feature map of the $(l-1)^{\text{th}}$ layer; y_l is the feature map of the l^{th} layer; M is the number of convolution kernels of the l^{th} layer.

By means of this multi-layer adaptive convolution mechanism, the adaptive convolution network can efficiently extract spatio-temporal characteristics of the dance motions from various scales and levels, so improving the recognition capacity of the model for complex dance movements.

4.3 Transfer learning module

Following feature extraction from the adaptive convolutional network, TL is applied to improve the model even more in dance movement identification. The lack of dance movement data allows TL to migrate the information acquired in the source domain to the target task, hence enhancing the generalising capacity of the model (Zhou et al., 2023).

Under the TL framework, the adaptive convolutional network learns generic features by training on the source domain data assuming source domain model is M_S and target domain model is M_T. TL moves these characteristics then to the target domain. The model of the source domain generates:

$$f_S(x_S) = y_S \tag{18}$$

where x_S is the source domain's input data; $f_S(\cdot)$ is the model trained in the source domain; y_S is the source domain's labels. By means of TL, the target domain model is tailored to inherit the convolutional layer weights of the source domain and implements particular modifications to the target task. The model output of the target domain can be stated assuming x_T as the data of the target domain as:

$$f_T(x_T) = f_S(x_S) + \Delta f(x_T) \tag{19}$$

where $\Delta f(x_T)$ is the component tuned on the target domain; x_T is the input of the target domain; $f_T(\cdot)$ is the output of the target domain model. In this sense, TL can not only speed up the training process of the target task but also enhance the recognition accuracy of the model using less target domain data.

4.4 Action recognition module

Action recognition is fundamentally about extracting and categorising the characteristics of every input frame using the model. Assuming that the input video frame is x_t, following processing by adaptive convolutional network, where t is the number of the frame and f_t is the feature representation of the frame picture. The model's last result is a categorisation one that shows the frame image falls into a certain dance movement category. Every frame's classification can be stated with the following equation:

$$\hat{y}_t = \text{softmax}\left(W \cdot f_t + b\right) \tag{20}$$

where W is the weight matrix of the classification layer; b is the bias term; softmax(·) is the activation function applied to translate model output into a probability distribution showing the expected probability of every category.

Using a time series modelling technique, the movement recognition model captures the dynamics of dance motions in the time dimension (Switonski et al., 2019). Temporal dependency between frames is modelled in this module using LSTM. The model changes the current hidden state h_t depending on the hidden state h_{t-1} at the previous moment and the current input x_t assuming that the hidden state at the current moment is h_t. One can articulate this process by means of the following equation:

$$h_t = \tanh\left(W_h \cdot h_{t-1} + W_x \cdot x_t + b_h\right) \tag{21}$$

where b_h is the bias term; tanh(·) is the activation function; W_h and W_x are weight matrices with regard to the hidden state and input features respectively. Maximising the probability values of the output categories helps the movement recognition model to precisely identify the dance movement categories depending on spatial characteristics and time series.

Combining TL and adaptive convolutional networks helps the model in the dance movement detection challenge to migrate knowledge between several dance forms with efficiency. For instance, the generalised movement properties of ballet data are applied to street or folk dances using TL, hence enhancing the generalisation capacity and accuracy of the model.

4.5 Optimisation and fine-tuning module

One can maximise the accuracy of the pre-trained ACN-TL-DAR model by adjusting its parameters in dance motion recognition. Fine-tuning aims to modify some network parameters such that they fit the new purpose.

Following model training, the module on optimisation and fine-tuning helps the model to perform on particular tasks. A gradient descent method with an update rule helps to optimise the fine-tuning process:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta L_{finetune}(\theta) \tag{22}$$

where $\nabla\theta L_{finetune}(\theta)$ is the gradient of the loss function; θ_t is a model parameter; η is the learning rate.

A reduced learning rate is utilised during the fine-tuning process to prevent damaging the knowledge of the pre-trained model, therefore allowing the model to reach ideal performance on the goal task (Shi and Lipani, 2023).

4.6 Evaluation and feedback module

The evaluation and feedback module is applied to verify the performance of the model in dance movement recognition following training and optimisation. Three main criteria help to guide the evaluation.

The ratio of correct recognition by the model on all movement categories with the following formula defines movement classification accuracy:

$$\textit{Action classification accuracy} = \frac{\sum_{i=1}^{N} \mathbb{I}\left(y_i = \hat{y}_i\right)}{N} \tag{23}$$

where N is the overall sample count; y_i is the actual label; $\hat{y}_i$ is the label the model forecasts. On several action categories, this statistic can show the general model performance.

With this method, one can determine timing consistency by verifying if the model can sustain consistency across consecutive video frames and appropriately identify intricate action variations (Sheng et al., 2021):

$$\text{Temporal consistency} = \frac{1}{N-1} \sum_{t=1}^{N-1} \mathbb{I}\left(\hat{y}_t = \hat{y}_{t+1}\right) \tag{24}$$

where N is the total number of frames; $\hat{y}_t$ and $\hat{y}_{t+1}$ correspondingly indicate the prediction categories of consecutive frames. Especially the transition and smoothness of the motions, this evaluation criterion shows the stability of the model throughout consecutive time intervals.

Furthermore, computed in dance datasets with significant sample imbalances is a category balance assessment of the model (Thabtah et al., 2020). This statistic ensures that, given the volume of data, the model's predictions balance between categories and prevents some categories from being biased in favour of the model. Computing the weighted accuracy of every category will help one to gauge this:

$$\text{Category balance} = \frac{\sum_{i=1}^{C} w_i \times \text{Precision}_i}{\sum_{i=1}^{C} w_i} \tag{25}$$

where Precision_i is the movement's accuracy in category i; C is the number of movement categories; w_i is the category's sample weight.

These evaluation criteria allow a thorough assessment of the performance of the model in the dance movement identification challenge, therefore offering vital input for later model enhancement.

In order to fully evaluate the performance of the model, we also recorded the training time, memory consumption, and feasibility of real-time deployment. During the training phase, the model completes training on a server equipped with NVIDIA RTX 3090 GPUs with a total training time of 48 hours. In terms of memory consumption, the model consumes about 16 GB of GPU memory on average during training. For real-time deployment, we tested the inference time of the model on a single GPU with an average inference time of 30 ms per frame, which indicates that the model is capable of real-time processing and suitable for deployment in real applications.

The data preprocessing module refines input data; the feature extraction module uses adaptive convolutional networks to extract spatio-temporal features; the transfer learning module borrows knowledge from the source task; the action recognition module classifies features; and the model optimisation and fine-tuning module improves accuracy.

5 Experimental results and analyses

5.1 Datasets

Two typical datasets, UCF101 and CMU MoCap, were used for this work in order to validate the efficiency of the ACN-TL-DAR model. These two datasets not only cover dance motions but also include various kinds of complex movements suited for validation of the model in the recognition of dance movements of realistic subjects.

During the data collection process, we followed strict ethical guidelines to ensure that all participants provided explicit informed consent. To protect the privacy of participants, we anonymised all motion capture data, removing any information that could identify individuals. In addition, we paid particular attention to cultural sensitivity in the use of the data to ensure that the data was not used in a way that would cause harm or discomfort to any cultural group. Through these measures, we ensured ethical handling of the data and protection of participants' rights and interests.

Table 1 shows the details of both datasets.

There are 101 movement categories in the UCF101 dataset, spanning modern dance, street dance, etc. among other dance movements. Tasks involving video-based motion recognition fit it.

There are a lot of dance movement capture data in the CMU MoCap dataset, which is especially fit for high-precision dance movement identification tasks; fine human motion capture data in the dataset is mostly used for movement analysis and posture estimate.

These two datasets offer different action data for this work, so richly supporting model training at both video and motion capture levels, respectively, so verifying the performance of the model in complicated dance action recognition.

To ensure the quality and consistency of the data, we adopted the following annotation strategy: the annotation of all dance movements was done by a professional dance instructor and a data annotation team. The annotation team first analysed the videos frame by frame, identified the start and end points of each dance movement, and annotated them into the corresponding movement categories. To ensure the accuracy of the annotation, the results of each video are reviewed and proofread by at least two experts. In the case of disagreement between experts, it will be resolved by discussion and reference to more video frames.

In the frame extraction process, we use a key frame extraction method based on optical flow. Optical flow method can detect the motion information of objects in the video, which helps us to identify frames that contain important action changes. By calculating the optical flow differences between neighbouring frames, we select those frames with significant motion changes as key frames to reduce redundancy and improve processing efficiency.

The preprocessing process includes steps such as image denoising and normalisation. Image denoising uses Gaussian blurring algorithm to smooth out the noise in the image by convolution operation. The normalisation process compresses the pixel values of the image into the range [0, 1] to ensure consistency and quality of the input data. With these preprocessing steps, the model can handle image data under different lighting and contrast conditions more consistently.

5.2 Comparison of dance movement recognition effects

Two experiments aiming at assessing the ACN-TL-DAR model's performance are carried out in this work. Experiment 1 is a comparison experiment, in which the performance of the model is assessed on several criteria by means of other often used action recognition models. Experiment 2 is an ablation experiment, in which important modules are removed from the model to evaluate the contribution of every module on the performance. These two tests complement one another and jointly assist to assess the performance of the ACN-TL-DAR model.

Table 1 Information on UCF101 and CMU MoCap

Dataset name	*Type*	*Number of action categories*	*Sample size*	*Action types*
UCF101	Video dataset	101	13,320	Includes dance actions
CMU MoCap	Motion capture dataset	Not explicitly defined	Not explicitly defined	Dance, human motion

Using a cross-valuation technique to guarantee the dependability of the experimental data, experiment 1 intends to evaluate the performance of the ACN-TL-DAR model against existing fused deep learning models. The ACN-TL-DAR suggested in this work is model 1; other models for experimental comparison consist:

- CNN + RNN is model 2: CNN extracts spatial features and RNN captures temporal information, so fitting for simulating continuous actions by combining the benefits of CNN and RNN.
- Model 3 is CNN plus TL: using TL, pre-training CNN (e.g., ResNet) extracts features and used to dance movement detection, so improving the performance of small datasets particularly in cases of inadequate data (Pramono et al., 2021).
- Appropriate for many dancing moves, model 4 is ACN + CNN: integrating adaptive convolutional network with CNN, adaptive convolutional network dynamically changes the convolution kernel to improve the feature extracting.
- Good in capturing long-term dependencies and overcomes the issue of long-term and short-term dependencies of typical RNNs, model 5 is TCN: temporal convolutional networks describe time series through convolutional layers.
- Appropriate for video classification and action identification problems, model 6 is 3D-CNN: expands conventional CNN to simultaneously extract spatio-temporal data via 3D convolution (Guo et al., 2019).
- Combining LSTM and CNN, model 7 is LSTM + CNN, LSTM processes temporal input and CNN extracts spatial data to enhance the accuracy of multimodal action detection.

In order to ensure the quality and reliability of the data, we have adopted a rigorous labelling validation and generation process. The labelling of all dance movements is done by a professional dance instructor and an experienced data labelling team. The annotation team first analyses the video frame by frame, identifies the start and end points of each dance movement, and labels them into the corresponding movement categories. To ensure the accuracy of the annotation, the results of each video are reviewed and proofread by at least two experts. In case of disagreement between experts, it will be resolved through discussion and reference to more video frames to ensure that the final annotation results are accurate.

For data enhancement, we use random cropping, random rotation, colour adjustment and time dithering to increase the diversity and robustness of the data. Random cropping can simulate different viewing angles and shooting ranges; random rotation has an angle range of [–10°, 10°] to simulate different shooting angles; colour adjustment simulates different lighting conditions by randomly adjusting the brightness, contrast and saturation of the image; and temporal dithering slightly randomises the temporal order of the video frames to increase the diversity of the temporal dimension.

In terms of data cleaning, we took measures to remove noise, remove outliers, remove inconsistent labelling and remove duplicate data. Gaussian blurring algorithm is used to remove noise from the images to ensure image clarity; frames with motion blur are removed by calculating optical flow differences to ensure the quality of each frame; video clips that are inconsistently labelled or contain a lot of noise are removed to improve the overall quality of the data; and duplicates or highly similar clips are removed by calculating similarity of the video clips to reduce data redundancy.

Through these data enhancement and cleaning steps, we ensure the high quality and reliability of the dataset, providing a solid foundation for model training and validation.

Figure 4 displays the experimental findings.

The ACN-TL-DAR model beats like models on both datasets. On the UCF101 data, ACN-TL-DAR boasts 93.2% movement classification accuracy, 91.5% temporal consistency, and 90.8% category balance. These results show that the suggested method precisely and consistently identifies complicated dancing movements and is suitable to solve inter-category balancing. While their time consistency and category balance are poor, 3D-CNN and CNN+RNN have acceptable classification accuracy.

Results of experiments on the CMU MoCap dataset validate the advantages of ACN-TL-DAR. Its action categorisation accuracy is 95.1%. In timing consistency (93.4%) and category balance (92.7%), ACN-TL-DAR is strong and flexible. These findings suggest that the suggested model might effectively handle distributional imbalance between categories and better represent dance motion temporal characteristics.

On both datasets, ACN-TL-DAR excels particularly in dance movement detection. This shows the success of combining adaptive convolutional networks with TL as well as the efficiency and feasibility of the model in challenging dynamic settings. Particularly in pragmatic applications, these comparative tests verify the movement recognition capacity of ACN-TL-DAR.

Figure 4 Results of the comparison of dance movement recognition effects (see online version for colours)

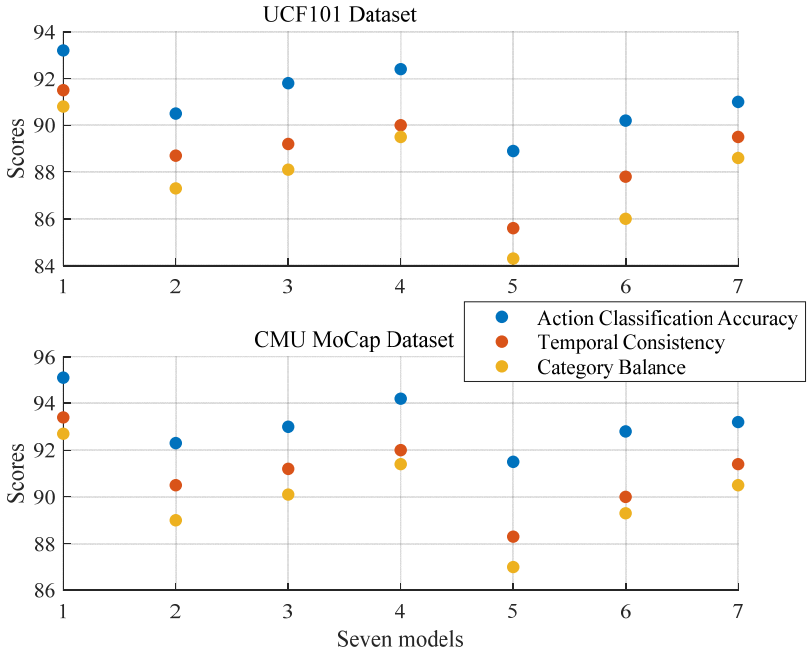
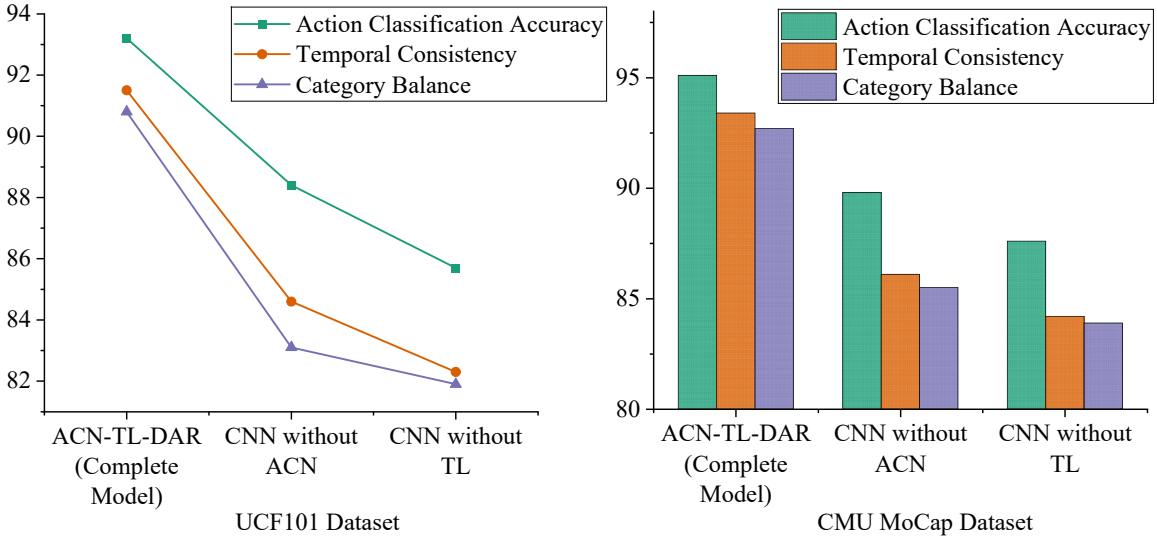


Figure 5 Results of the comparison of dance movement recognition performance (see online version for colours)



5.3 Effect of different modules on recognition performance

This experiment will examine the change in model performance by subjecting the ACN-TL-DAR model to a sequential ablation deleting the adaptive convolutional network and TL sections accordingly. This allows one to estimate and evaluate the contributions of the adaptive convolutional network and the TL in line with the whole model (ACN-TL-DAR).

Eliminating the adaptive convolutional network: maintaining the TL component unaltered, replace the adaptive convolutional network with a standard CNN.

Eliminating TL: CNNs feature extraction and the entire model is trained from scratch without utilising TL.

Keep the TL module and adaptive convolutional network for last performance assessment.

Figure 5 shows the experimental outcomes.

Especially in the three indexes of action classification accuracy, temporal consistency and category balance, which are all better than that of the model after removing the adaptive convolutional network and TL, the ACN-TL-DAR full model shows optimal performance from the experimental results on both UCF101 and CMU MoCap datasets. Specifically, the entire model obtains an action classification accuracy of 93.2% on the UCF101 dataset, whereas the model with the removal of adaptive convolutional network has a loss in accuracy of roughly 4.8%; and the model with the removal of TL has a decrease in accuracy of roughly 7.5%. This performance difference implies that adaptive convolutional networks and TL can

efficiently increase feature extraction and model generalisation as well as model accuracy, particularly with regard to difficult dancing motions.

The movement classification accuracy of the whole ACN-TL-DAR model is 95.1%, which is 5.3% and 7.5% higher than that of the model with the removal of adaptive convolutional network (89.8%), and the model with the removal of TL (87.6%), respectively using the CMU MoCap dataset. Furthermore, displaying notable benefits in temporal consistency and category balance is the whole model. Further underlining the relevance of the adaptive convolutional network and TL in enhancing the model in terms of timing processing and category balance, their removal reduced the timing consistency and category balance of the model.

The findings of the ablation studies confirm the efficacy of the ACN-TL-DAR model, particularly concerning temporal characteristics and category imbalance. The incorporation of adaptive convolutional networks and TL significantly enhances the model's overall performance. This underscores the potential of integrating deep learning techniques with TL for identifying dance movements in realistic subjects, thereby addressing the variety and challenges present in complex environments.

6 Conclusions

This work presents an adaptive convolutional network and TL-based dance action recognition model ACN-TL-DAR for realistic subjects, together assessed on UCF101 and CMU MoCap datasets. Comparative and ablation studies confirm the major benefits of the model in terms of action categorisation accuracy, temporal consistency and category balance. Particularly in the complicated dance movement detection test, the ACN-TL-DAR model shows great resilience and accuracy, therefore confirming the value of adaptive convolutional networks and TLs in this field.

Despite the impressive performance of the model proposed in this work across various tests, there are still significant limitations. Firstly, the training procedure of the model relies heavily on a substantial amount of labelled data, which is particularly difficult to obtain, especially when it comes to realistic dancing data. Secondly, the model may not perform as expected in certain specific scenarios, particularly when there are fewer movement categories or an unequal distribution of categories, due to feature variances present in different datasets. Additionally, when training on large-scale datasets, there is a need for more efficient computational resources and optimisation techniques. Therefore, there remains an opportunity to enhance the model in terms of computational complexity and training time.

Future lines of study can concentrate on the following: first, investigating semi-supervised or unsupervised learning methods based on a limited amount of labelled data to lower the dependability on labelled data (Qi and Luo, 2020); second, combining multimodal data (e.g., video, audio, and motion-capture data) to further improve the generalisation ability and recognition accuracy; finally, optimising the model's computational efficiency can be considered to develop more efficient algorithms to adapt to the training needs of large-scale datasets. These developments should help to increase the practical relevance of realistic dance movement recognition technology and support higher level of research in this domain.

Declarations

This work is supported by the key scientific research project of Hunan Provincial Department of Education named: Research on Red Dance Creation and its Contemporary Value under the Background of New Era (No. 24A0597).

The author declares that she has no conflicts of interest.

References

Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M.A., Al-Amidie, M. and Farhan, L. (2021) 'Review of deep learning: concepts, CNN architectures, challenges, applications, future directions', *Journal of Big Data*, Vol. 8, pp.1–74.

Andersson, P., Nilsson, J., Akenine-Möller, T., Oskarsson, M., Åström, K. and Fairchild, M.D. (2020) 'FLIP: a difference evaluator for alternating images', *Proc. ACM Comput. Graph. Interact. Tech.*, Vol. 3, No. 2, pp.15:1–15:23.

Borji, A., Seifi, A. and Hejazi, T.H. (2023) 'An efficient method for detection of Alzheimer's disease using high-dimensional PET scan images', *Intelligent Decision Technologies*, Vol. 17, No. 3, pp.729–749.

Channoufi, I., Bourouis, S., Bouguila, N. and Hamrouni, K. (2018) 'Image and video denoising by combining unsupervised bounded generalized gaussian mixture modeling and spatial information', *Multimedia Tools and Applications*, Vol. 77, pp.25591–25606.

Chen, K., Zhang, D., Yao, L., Guo, B., Yu, Z. and Liu, Y. (2021) 'Deep learning for sensor-based human activity recognition: overview, challenges, and opportunities', *ACM Computing Surveys (CSUR)*, Vol. 54, No. 4, pp.1–40.

Dargan, S., Kumar, M., Ayyagari, M.R. and Kumar, G. (2020) 'A survey of deep learning and its applications: a new paradigm to machine learning', *Archives of Computational Methods in Engineering*, Vol. 27, pp.1071–1092.

Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C-M. and Chen, W. (2023) 'Parameter-efficient fine-tuning of large-scale pre-trained language models', *Nature Machine Intelligence*, Vol. 5, No. 3, pp.220–235.

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.

Guo, S., Lin, Y., Li, S., Chen, Z. and Wan, H. (2019) 'Deep spatial-temporal 3D convolutional neural networks for traffic data forecasting', *IEEE Transactions on Intelligent Transportation Systems*, Vol. 20, No. 10, pp.3913–3926.

Kong, Y. and Fu, Y. (2022) 'Human action recognition and prediction: a survey', *International Journal of Computer Vision*, Vol. 130, No. 5, pp.1366–1401.

Li, F., Zhou, H., Wang, Z. and Wu, X. (2020) 'ADDCNN: an attention-based deep dilated convolutional neural network for seismic facies analysis with interpretable spatial-spectral maps', *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 59, No. 2, pp.1733–1744.

Li, H., Li, X., Su, L., Jin, D., Huang, J. and Huang, D. (2022) 'Deep spatio-temporal adaptive 3D convolutional neural networks for traffic flow prediction', *ACM Transactions on Intelligent Systems and Technology (TIST)*, Vol. 13, No. 2, pp.1–21.

Liu, S., Li, X., Zhai, Y., You, C., Zhu, Z., Fernandez-Granda, C. and Qu, Q. (2021a) 'Convolutional normalization: improving deep convolutional network robustness and training', *Advances in Neural Information Processing Systems*, Vol. 34, pp.28919–28928.

Liu, Y., Pu, H. and Sun, D.-W. (2021b) 'Efficient extraction of deep image features using convolutional neural network (CNN) for applications in detecting and analysing complex food matrices', *Trends in Food Science & Technology*, Vol. 113, pp.193–204.

Newell, K.M. (2020) 'What are fundamental motor skills and what is fundamental about them?', *Journal of Motor Learning and Development*, Vol. 8, No. 2, pp.280–314.

Panadeiro, V., Rodriguez, A., Henry, J., Wlodkowic, D. and Andersson, M. (2021) 'A review of 28 free animal-tracking software applications: current features and limitations', *Lab Animal*, Vol. 50, No. 9, pp.246–254.

Pramono, R.R.A., Chen, Y-T. and Fang, W-H. (2021) 'Spatial-temporal action localization with hierarchical self-attention', *IEEE Transactions on Multimedia*, Vol. 24, pp.625–639.

Qi, G-J. and Luo, J. (2020) 'Small data challenges in big data era: a survey of recent progress on unsupervised and semi-supervised methods', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 44, No. 4, pp.2168–2187.

Sheng, B., Li, P., Ali, R. and Chen, C.P. (2021) 'Improving video temporal consistency via broad learning system', *IEEE Transactions on Cybernetics*, Vol. 52, No. 7, pp.6662–6675.

Shi, Z. and Lipani, A. (2023) 'Don't stop pretraining? Make prompt-based fine-tuning powerful learner', *Advances in Neural Information Processing Systems*, Vol. 36, pp.5827–5849.

Switonski, A., Josinski, H. and Wojciechowski, K. (2019) 'Dynamic time warping in classification and selection of motion capture data', *Multidimensional Systems and Signal Processing*, Vol. 30, pp.1437–1468.

Thabtah, F., Hammoud, S., Kamalov, F. and Gonsalves, A. (2020) 'Data imbalance in classification: experimental evaluation', *Information Sciences*, Vol. 513, pp.429–441.

Wang, Y., Yan, J., Yang, Z., Qi, Z., Wang, J. and Geng, Y. (2022) 'Gas-insulated switchgear insulation defect diagnosis via a novel domain adaptive graph convolutional network', *IEEE Transactions on Instrumentation and Measurement*, Vol. 71, pp.1–10.

Zhang, H., Le, Z., Shao, Z., Xu, H. and Ma, J. (2021) 'MFF-GAN: an unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion', *Information Fusion*, Vol. 66, pp.40–53.

Zhou, Q., Li, M., Zeng, Q., Aristidou, A., Zhang, X., Chen, L. and Tu, C. (2023) 'Let's all dance: enhancing amateur dance motions', *Computational Visual Media*, Vol. 9, No. 3, pp.531–550.