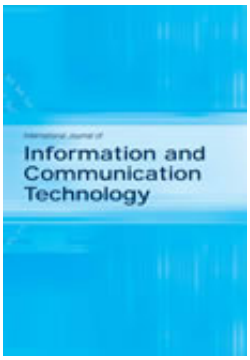
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

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

Dynamic segmentation algorithm of dance video image for non-legacy ethnic dance inheritance

Xiaoyu Zhang

DOI: 10.1504/IJICT.2025.10074932

Article History:

Received:	30 September 2025
Last revised:	22 October 2025
Accepted:	28 October 2025
Published online:	17 December 2025

Dynamic segmentation algorithm of dance video image for non-legacy ethnic dance inheritance

Xiaoyu Zhang

School of Fashion,
Henan University of Engineering,
No. 1 Xianghe Road, Longhu Town,
Xinzheng City, Zhengzhou, 450000, Henan Province, China
Email: 17513132182@163.com

Abstract: Traditional manual arrangement struggles to efficiently handle complex backgrounds, frequent movements and variable lighting in dance videos; high-precision automatic techniques are urgently needed for digital analysis and protection. In this paper, a large-scale dance video dataset covering multi-ethnic, multi-scene and multi-illumination conditions is constructed, and the preliminary extraction of foreground region is realised by using the motion detection module combining frame difference method and optical flow method. Then, based on the improved U-Net structure, multi-scale feature fusion and attention mechanism are designed to enhance the segmentation ability of clothing and limb details, and the joint loss of Dice and cross entropy is used to improve the boundary accuracy. Experimental results show that the proposed method is better than U-Net and DeepLabv3+ in terms of IoU, Dice, precision, recall, F1-score, etc., and shows stronger robustness and near real-time processing speed in complex scenes.

Keywords: intangible cultural heritage; ethnic dance; video image processing; dynamic segmentation algorithm.

Reference to this paper should be made as follows: Zhang, X. (2025) 'Dynamic segmentation algorithm of dance video image for non-legacy ethnic dance inheritance', *Int. J. Information and Communication Technology*, Vol. 26, No. 47, pp.106–127.

Biographical notes: Xiaoyu Zhang is affiliated with the School of Fashion, Henan Institute of Engineering, Zhengzhou, Henan Province. He holds a Bachelor's degree (2007–2011) from Huaihua University and Master's degree (2011–2013) from Luhansk Taras Shevchenko National University in Ukraine.

1 Introduction

The segmentation of ethnic dance videos presents not only technical but also cultural challenges that differ significantly from general video segmentation tasks. Technically, ethnic dances involve rapid, non-rigid body movements, complex interactions between multiple dancers, and frequent occlusions caused by traditional costumes and accessories. These characteristics complicate motion detection and boundary recognition. Culturally, dance movements often encode symbolic gestures and heritage-specific expressions that are vital to intangible cultural preservation. A segmentation algorithm must therefore

capture these features without compromising their authenticity. Existing methods in computer vision have mainly focused on urban scenes, daily actions, or medical imaging, leaving a gap in dealing with the aesthetic and symbolic dimensions of ethnic performance. This study addresses this gap by proposing a dynamic segmentation framework that integrates cultural context into algorithmic design.

The preservation of intangible cultural heritage, especially ethnic dance, has become increasingly urgent as modernisation accelerates and traditional art forms face the risk of fragmentation. Beyond cultural aesthetics, dance embodies social memory, ritual identity, and intergenerational knowledge. However, the digitisation of ethnic dances introduces complex technical and cultural challenges. The intricate costumes, collective choreography, and symbolic gestures inherent to many ethnic performances pose difficulties for existing video segmentation algorithms. These challenges underline a cultural urgency: how to ensure that digital technologies not only record movement but also preserve meaning. This study responds to that urgency by developing an intelligent segmentation framework that bridges cultural preservation and algorithmic innovation, transforming the act of dance analysis into a form of cultural continuity.

The introduction outlines the technical basis of dance video segmentation but also emphasises its broader cultural and technological implications. Applying segmentation algorithms to intangible cultural heritage contexts is not only a computational challenge but also a reflection of how technology mediates cultural expression. In ethnic dance, movements embody collective memory, identity, and ritual significance. The accurate segmentation of such performances goes beyond boundary detection – it preserves cultural semantics embedded in motion. Expanding this perspective connects computer vision research with cultural sustainability, showing that algorithmic innovation can actively participate in the documentation and revitalisation of living traditions.

Across the world, digitising ethnic dance heritage remains a formidable challenge. For example, the multilayered costumes of Tibetan and Mongolian dances often produce reflective surfaces and complex shadows, while African or Polynesian group choreographies involve overlapping dancers and rapid collective formations that hinder reliable segmentation. Such visual intricacies require dynamic algorithms that can recognise and separate culturally meaningful motion patterns from environmental noise. Despite progress in computer vision and motion analysis, most prior research has focused on urban actions, sports, or biomedical signals, leaving a significant gap in addressing the cultural and performative complexity of traditional dances. Understanding these global challenges highlights the necessity of developing segmentation models sensitive to both motion precision and cultural semantics. Although the literature on motion segmentation and computer vision is extensive, it remains largely descriptive when applied to non-rigid and complex dance movements. Earlier models, such as frame-difference and optical-flow-based methods, struggle to distinguish overlapping dancers and costume-induced noise. Deep learning approaches improve precision but often neglect the temporal continuity and expressive semantics of ethnic choreography. The absence of critical reflection on these limitations weakens their applicability to cultural heritage data. A key challenge lies in developing models that can adapt to the fluidity, complexity, and symbolic nature of dance while maintaining algorithmic efficiency and interpretability. This study builds upon that gap by combining motion analysis with a heritage-sensitive design perspective.

The research on national dance of intangible cultural heritage presents a systematic development trend in education, cultural inheritance, data analysis and intelligent algorithm. In the field of education, Wang (2025) focuses on the relationship between dance education and students' mental health under the background of national culture, pointing out that school dance education can play a role in promoting students' psychological adjustment in the inheritance of national culture, and emphasises the important position of dance in the construction of emotional expression and identity. Mao et al. (2025) studied Tujia dance therapy from the perspective of cultural anthropology and health science, and put forward that national dance has specific cultural symbols and action characteristics, which shows unique value in emotional regulation and physical and mental health. This kind of research provides a clear theoretical support for the cultural attribute and educational function of national dance.

In the aspect of digital analysis of cultural resources, Luo et al. (2025) used big data analysis method to explore the experience attribute of intangible cultural heritage, and thought that digital technology was helpful to reveal the potential structure of cultural experience and provide a data basis for digital modelling and dissemination of dance intangible cultural heritage. Gao et al. (2024) put forward a data-driven signal automatic segmentation and feature fusion method for EEG emotion recognition, emphasising the advantages of automatic segmentation technology in feature extraction of complex dynamic signals, which has reference significance for the segmentation of multi-time sequence and multi-dimensional action features in dance videos. Wu et al. (2024) analysed the spatial distribution and influencing factors of intangible cultural heritage in Yunnan, Guangxi and Guizhou, and pointed out that geographical environment and human factors play an important role in cultural distribution and protection. This spatial perspective provides a basis for the regional analysis and inheritance strategy of national dance data.

At the level of algorithm and model, Sjobeck et al. (2024) put forward a pairwise approximate spatio-temporal symmetry algorithm for time series segmentation, thinking that this method can improve the segmentation accuracy while maintaining the consistency of spatio-temporal structure. Ma et al. (2024) introduced multi-feature coordinate learning method in medical image segmentation, emphasising that feature fusion can improve the accuracy of complex structure recognition, which is highly consistent with the requirements of capturing edges and details in dance video segmentation. Fan (2024) proposed a framework of description and knowledge fusion of intangible resources based on associated data, and thought that the unified expression of multi-source heterogeneous data was the core of intangible digitalisation. Song et al., (2023) proposed a path optimisation method for dynamic time-varying networks, revealing the importance of time series characteristics in dynamic system analysis. Su et al. (2023) identified the potential group structure through the spatial dynamic panel, and proposed a multi-dimensional spatio-temporal data modelling method, which provided an idea for the structuring of national dance data in complex scenes. Wang et al. (2023) analysed the spatial distribution characteristics and influencing factors of intangible cultural heritage in the Yangtze River basin, emphasising the role of regional cultural ecology in non-genetic inheritance. Fan et al. (2023) studied the knowledge organisation of intangible spatio-temporal data from the perspective of digital humanities, and pointed out that structuring and semantic processing are the key to digital protection and dissemination. Previous studies have made significant progress in motion segmentation and pose estimation; however, most approaches lack a critical synthesis of

how these methods perform under culturally complex conditions. Traditional segmentation algorithms often fail to deal with non-rigid and fast-changing movements typical of ethnic dances. Furthermore, the intricate patterns and multilayered textures of traditional costumes introduce additional noise that conventional edge-based or region-based methods cannot effectively handle. From a cultural anthropology perspective, prior research has rarely integrated cultural semantics and expressive patterns into technical modelling, resulting in outputs that are visually accurate but semantically incomplete. The present research builds upon these insights by proposing an integrated framework that merges motion analysis with cultural feature preservation.

The primary objectives of this research are threefold:

1 to construct a dynamic segmentation framework tailored to the characteristics of ethnic dance videos

2 to develop a motion region detection module that enhances the precision of non-rigid body segmentation under complex visual conditions

3 to evaluate the proposed model's performance in terms of accuracy, robustness, and applicability to digital heritage preservation.

These objectives collectively guide the overall research design and provide a coherent structure for subsequent methodological and analytical sections. In this process, the key problems to be solved are as follows:

1 how to separate dancers from the background accurately under strong background interference and illumination changes

2 how to ensure the stability and real-time performance of the algorithm under the conditions of low frame rate and different shooting equipment

3 how to improve the processing efficiency of the algorithm for large-scale video data on the premise of ensuring accuracy.

The solution of these problems will lay a foundation for the digital archiving and intelligent dissemination of national dances.

To solve the above problems, computer vision, deep learning and video processing technology are comprehensively used to build a video dynamic segmentation technology system for non-legacy national dances. On the data level, a high-quality video dataset covering many kinds of folk dances is established, and the accuracy of the model input data is improved by using time synchronisation frame extraction, denoising and abnormal frame cleaning techniques. At the algorithm level, the motion detection strategy combining frame difference method and optical flow method is adopted to realise the preliminary extraction of motion subjects in dance videos. Based on the improved U-Net architecture, a multi-scale attention fusion module is embedded to enhance the feature extraction ability. In the training and verification stage, the joint optimisation strategy of Dice loss and cross entropy loss is introduced to improve the segmentation effect of the model on dancers' edges and detail areas. The generalisation ability and robustness of the model are evaluated through the comparative experiments of multi-scene, multi-dance and multi-frame rate. On the application level, this study combines the experimental results with the actual needs of digital protection of national dances, and verifies the applicability of the algorithm in complex festival activities, traditional performance

recording and digital display platforms by analysing the performance of different dances and scenes. The overall research method takes into account both theoretical innovation and application, and balances segmentation accuracy, adaptability and processing efficiency.

2 Materials and methods

2.1 Data collection and sample selection

2.1.1 Data sources and collection methods

To ensure the representativeness and diversity of the constructed dance video dataset of non-legacy ethnic groups, the data sources are divided into three categories: official archives resources, field collected videos and social public video materials. First of all, the official archives resources mainly come from the intangible cultural heritage digital engineering achievement database of the ministry of culture and tourism and local cultural centres, including the national and provincial intangible cultural heritage dance project videos recorded between 2020 and 2024 (Fan et al., 2022). Most of these videos were shot by professional camera teams in stage performances and festivals, with high quality and stable lens switching characteristics. About 250 high-definition videos were collected, most of which had a resolution of 1080p and a few of which were 4K.

The field collection was carried out in Sichuan, Guizhou, Xizang, Inner Mongolia and other regions during 2023–2024, covering many dance types, such as Miao Lusheng dance, Dong Da Ge dance, Tibetan pot dance and Mongolian Andai dance. Multi-camera synchronous shooting system is used, including three Sony PXW-FX6 cameras and DJI Mavic 3 aerial drone, which is used to capture the performance process from the front, oblique side and top view. The frame rate is uniformly set at 30 fps, and the duration is 3–10 minutes, and 120 self-built data are collected, with a total duration of more than 15 hours.

The third category is video materials from the public, including online public videos of various folk festivals and non-legacy performances (Bilibili, CCTV videos, etc.), which are included in the dataset after copyright confirmation. This part of the material has a variety of image quality, such as large illumination change, complex background and crowd occlusion, which is suitable for verifying the robustness of the algorithm. A total of 80 videos are collected. To sum up, the final dataset contains 450 videos, with a total duration of about 52 hours and a total number of frames exceeding 580,000, covering multi-ethnic, multi-scene and multi-shooting equipment conditions, which can comprehensively reflect the diversity and complexity of non-legacy national dances in the actual inheritance process.

2.1.2 Sample selection and description

Sample selection takes into account the diversity of national dance movements, costumes and scenes. In this study, stratified sampling is carried out according to dance types, performance scenes, shooting methods and image quality levels to ensure the balance of training and testing data. In terms of dance types, four representative types are mainly selected:

1 Tibetan pot dance (frequent rotation and elegant costumes)

2 Lusheng dance of Miao nationality (mainly group dance with strong action rhythm)

3 Dong people's big songs and dances (the combination of singing and dancing, complicated costumes)

4 Mongolian Andai dance (with large amplitude and stable movement rhythm).

Each kind of dance includes about 50 videos of stage performance scene and outdoor festival scene, which is used to test the adaptability of the algorithm under different background complexity.

Shooting video includes not only high-quality materials shot synchronously by three cameras, but also unstable pictures shot by handheld devices, which is especially common in folk festivals. The image quality level covers 720p, 1080p and 4K, and the materials with different resolutions can be used to analyse the performance differences of the segmentation algorithm under the change of resolution. The whole dataset is divided into training set (315 segments), verification set (90 segments) and test set (45 segments) according to the ratio of 7:2:1, so as to ensure that the distribution of all kinds of dances, scenes and image quality levels in each subset is basically the same. In order to prevent data leakage caused by the intersection of scenes and dance types, the video of the same performance will not appear in the training set and test set at the same time. In addition, for the group dance scene with a large number of dancers, meta-information such as the number of dancers, the size of stage space and the background type are also recorded in the sample description, which is used to analyse the relationship between segmentation accuracy and scene characteristics.

2.1.3 Data preprocessing

Before the formal training, all the collected videos were preprocessed in a unified way, including time synchronisation frame extraction, denoising, resolution unification and abnormal frame detection (as shown in Table 1). All videos are frame extracted according to the original frame rate, and the resolution is uniformly adjusted to 1,920 × 1,080 to ensure the consistent input size. Secondly, the time domain filtering method based on wavelet transform is used to remove the high-frequency noise in the video and stabilise the picture jitter caused by handheld devices.

For some videos with significant illumination changes, adaptive histogram equalisation (CLAHE) is used to improve the contrast and reduce the influence of ambient light. Abnormal frame detection is mainly realised by time continuity detection and edge energy change analysis, and empty frames, repeated frames and seriously blurred frames are deleted. After preprocessing, the overall quality and stability of video frames are significantly improved, which provides high-quality input for subsequent motion detection and segmentation network training.

It can be seen from Table 1 that about 25,000 invalid frames were eliminated in the preprocessing, accounting for about 4.3%, which significantly improved the data quality and ensured the consistency and availability of training data.

Table 1 Statistics of video data before and after preprocessing

Data source category	*Original frame number*	*Missing frames*	*Noise frame number*	*Preprocessed frame number*
Official file video	250,000	5,000	3,200	241,800
Self-collected video	220,000	4,300	2,700	213,000
Public video materials	110,000	6,100	3,900	100,000
Total	580,000	15,400	9,800	554,800

2.1.4 Data cleaning

After preprocessing, data cleaning and quality control are carried out, and the abnormal segments with discontinuous time axis or excessive difference between frames are deleted by using timestamp sequence and inter-frame correlation analysis to ensure the time continuity of dance movements (Zhao, 2022). Secondly, all video frames were manually sampled for quality inspection, and 20 frames were randomly selected from each video, which were manually marked and reviewed by two researchers with dance background, and videos with inconsistent labels or extremely low picture quality were deleted, and about 3% of the data were excluded. For videos from public sources, the background noise detection algorithm is adopted. By analysing the edge density and brightness change characteristics of the background area, the segments containing serious occlusion (such as audience), subtitle occlusion or obvious non-dance content are identified, and about 2,000 frames are further cleaned (Abbasimehr and Bahrini, 2022). To improve the labelling quality of datasets, semi-automatic labelling tools (labelme + optical flow assistance) are used to label the dancer's contour initially, and manual secondary correction is made to ensure the accuracy of segmentation labels (De Souza et al., 2022). The frame quality of the cleaned dataset is stable, the scene description is clear and the label accuracy is high, which lays a reliable data foundation for the subsequent training and evaluation of motion detection and depth segmentation network.

2.2 Model selection and construction

The methodological framework is informed not only by technical efficiency but also by principles derived from cultural heritage research. The selection of segmentation models and feature fusion strategies was guided by the need to preserve expressive authenticity in ethnic dance motion. For instance, the use of multi-scale feature fusion and temporal smoothing was motivated by the requirement to maintain visual coherence in complex group choreographies. Similarly, the choice of loss functions, such as Dice and perceptual loss, was influenced by the goal of balancing computational optimisation with the accurate rendering of culturally meaningful movement boundaries. This approach ensures that algorithmic design aligns with the ethical and aesthetic priorities of cultural preservation.

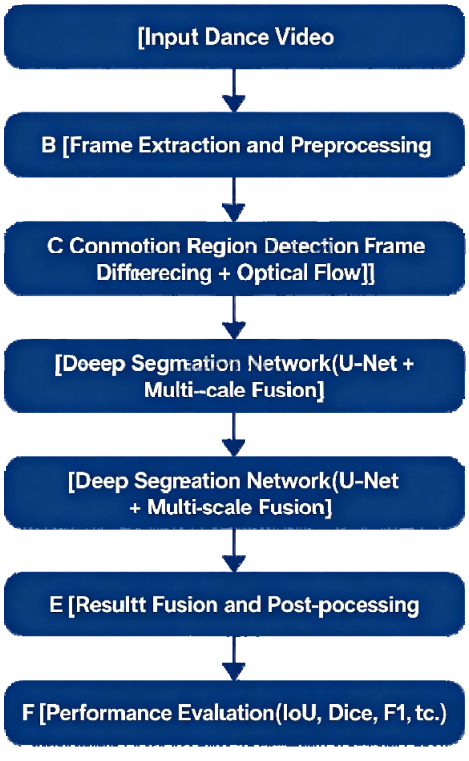
2.2.1 Overall process of dynamic segmentation of dance video images

Aiming at the characteristics of complex scene, great illumination change and high action frequency of dancers in the dance video of non-legacy ethnic groups, a set of overall processing flow of dynamic segmentation of dance video images is designed. The overall

framework includes five main stages: video preprocessing, motion region detection, feature extraction and segmentation, result fusion and optimisation, and performance evaluation (Figure 1). Pre-processing operations such as frame extraction, denoising, picture stabilisation and contrast enhancement are carried out on the input national dance video to ensure the temporal and spatial consistency of subsequent processing. Using the motion detection module combining frame difference method and optical flow method, the dancer's action area is initially located, which reduces the computational burden of the subsequent depth network.

In the feature extraction and segmentation stage, the improved U-Net network structure is adopted, combined with multi-scale feature fusion and attention mechanism, to enhance the model's ability to depict the elegant edge of dance costumes and the overlapping area of group dance. In the fusion stage, feature weight adjustment and morphological post-processing are used to improve the integrity and boundary accuracy of segmentation results. Finally, the segmentation results are evaluated by multi-dimensional indicators, including IoU, Dice coefficient, precision, recall and so on. The whole process gives consideration to both accuracy and real-time, and is suitable for video analysis tasks under the conditions of multi-ethnic dances and multi-shooting.

Figure 1 Overall flow of dynamic segmentation algorithm for dance video (see online version for colours)



2.2.2 Motion area detection method

In complex dance videos, the background area contains a large number of audiences, stage decorations and dynamic lighting effects, so it is necessary to screen out the static background in advance through motion area detection. By combining the frame difference method with the optical flow method, the dancer's movement area is detected in both time domain and spatial gradient level (Li and Gang, 2021).

1 Frame difference method based on the pixel difference of adjacent frames, the frame difference method is suitable for detecting high-speed and continuous motion areas in dance videos. Given two adjacent frames of images $I_t(x, y)$ and $I_{t-1}(x, y)$, the difference graph is defined as equation (1):

$$D_t(x, y) = \left| I_t(x, y) - I_{t-1}(x, y) \right| \tag{1}$$

(x, y) represents the spatial coordinates of the image, and $d_t(x, y)$ represents the grey level difference between the t^{th} frame and the $(t-1)^{th}$ frame at this position. By thresholding $D_t(x, y)$, the binary mask of the motion region can be obtained quickly. However, the frame difference method is sensitive to slow motion and illumination changes, and it is easy to produce missed detection and false detection.

2 In order to make up for the limitation of frame difference method, optical flow method is introduced to estimate the pixel motion vector field. According to the optical flow constraint equation, such as equation (2):

$$I_x u + I_y v + I_t = 0 \tag{2}$$

I_x and I_y are the horizontal and vertical gradients of the image, and I_t is the time gradient u and v represent the motion components of a pixel in the horizontal and vertical directions, respectively. The equation assumes that the pixel grey level remains constant during the movement. By calculating the gradient of adjacent frames, the optical flow field (u, v) can be obtained, and the areas with slow dance movements and light insensitivity can be further extracted (Kaholokula et al., 2021). The detection results of frame difference method and optical flow method are logically or-operated, and the noise is removed by connected domain analysis and Gaussian filtering to obtain a stable moving foreground region, which provides prior spatial information for subsequent depth segmentation.

2.2.3 Fusion of segmentation model structure and multi-scale features

Based on the classic U-Net structure, the depth segmentation model is improved according to the characteristics of national dance videos:

1 a multi-scale convolution layer is added at the coding end to capture the details of dancers' costumes and bodies at different scales

2 the attention mechanism is introduced at the decoding end, and the key areas are weighted and strengthened

3 improve the accuracy and integrity of segmentation boundary through multi-scale feature fusion module (Davis, 2020).

The core idea of multi-scale feature fusion is to fuse feature maps F_1, F_2, …, F_n from different scales by weighted summation, and the weights are adaptively learned by attention mechanism. The fusion equation (3) is as follows:

$$F_{fusion} = \sum_{i=1}^{n} \alpha_i \cdot F_i \tag{3}$$

F_i represents the feature map of the i^{th} scale, and α_i represents the corresponding attention weight, which satisfies $\sum_i \alpha_i = 1$. This method preserves the low-frequency information of dancers' overall movements and the high-frequency details of costumes and limb edges, and obtains more stable and accurate segmentation results under complex lighting and background conditions. To avoid information loss, the fusion module performs spatial alignment and scale normalisation before feature mosaic to ensure that features of different scales correspond strictly in the spatial dimension and improve the fusion effect (Munusamy and Murugesan, 2020).

2.2.4 Loss function and training strategy

To ensure that the model has higher segmentation accuracy in the edge and detail areas of dancers, the joint optimisation strategy of Dice loss and cross entropy loss is adopted. Dice loss is suitable for measuring the difference between the prediction results and the real mask in the spatial overlapping area, and can deal with the problem of uneven proportion of positive and negative samples in dance videos. Equation (4) is defined:

$$\mathcal{L}_{Dice} = 1 - \frac{2|P \cap G|}{|P| + |G|} \tag{4}$$

P represents the predicted foreground area, G represents the real marked area, and $|\cdot|$ represents the number of pixels. Improve the accuracy of overall pixel-level classification, and introduce binary cross entropy loss $\mathcal{L}_{CE}$. The final total loss function is defined as in equation (5).

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{Dice} + \lambda_2 \mathcal{L}_{CE} \tag{5}$$

λ_1 and λ_2 are the weight parameters of the two loss items, respectively. In this paper, $\lambda_1 = 0.6$ and $\lambda_2 = 0.4$ are taken and determined by verification set experiments.

Table 2 Model training parameter settings

Parameter name	*Value*
Optimiser	Adam
Initial learning rate	1×10^{-4}
Batch size	8
Number of training rounds	100
Loss function	Dice + cross entropy joint loss
Weight proportion	$\lambda_1 = 0.6$, $\lambda_2 = 0.4$
Data enhancement	Random cropping, rotation, illumination disturbance

As shown in Table 2, the training strategy adopts Adam optimiser, the initial learning rate is 1×10^{-4}, the batch size is 8 and the number of training rounds is 100. In order to prevent over-fitting, the early stop mechanism and data enhancement (random cropping, rotation and illumination disturbance) strategies are used to enhance the generalisation ability of the model.

2.3 Model evaluation and verification

2.3.1 Evaluation indicators

To evaluate the performance of the dynamic segmentation algorithm proposed in this paper in different ethnic dance videos, we use many classic evaluation indexes in the field of image segmentation, including cross-union ratio (IoU), Dice coefficient, precision, recall and F1-score. These indicators can reflect the overlapping degree, classification accuracy and overall consistency between segmentation results and real labelling from different angles. As shown in Table 3, IoU measures the degree of overlap between the predicted area and the real area. Dice coefficient emphasises the overlapping ratio between the predicted results and the real area, and has higher stability when the proportion of positive and negative samples is unbalanced (such as the disparity between dancers and background pixels). Precision measures how many pixels predicted as foreground are real foreground. Recall measures how many real prospects are successfully detected, and F1-score is the harmonic average of precision and recall. Through the joint use of these five indicators, the performance of the segmentation algorithm in different scenes and dances is comprehensively evaluated from multiple dimensions such as spatial overlap, boundary accuracy and pixel classification.

Table 3 Model performance evaluation indicators

Index	*Equation*	*Evaluation dimension*
IoU	$\text{IoU} = \frac{\lvert P \cap G \rvert}{\lvert P \cup G \rvert}$	$P \cap G$
Dice	$\text{Dice} = \frac{2\lvert P \cap G \rvert}{\lvert P \rvert + \lvert G \rvert}$	$P \cap G$
Precision	$\text{Precision} = \frac{TP}{TP + FP}$	Accuracy of positive class judgement
Recall	$\text{Recall} = \frac{TP}{TP + FN}$	Positive coverage rate
F1-score	$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$	Balance between accuracy and recall

2.3.2 Experimental design and cross-validation

Using systematic experimental design method and cross-validation strategy, the proposed dynamic segmentation algorithm of dance video is fully evaluated. The dataset is divided into training set, verification set and test set according to the ratio of 7:2:1, which ensures that the distribution of dance types, scenes and video quality in each subset is basically

the same and avoids bias. In the process of division, the clips of the same performance or the same video source are strictly isolated to prevent information leakage.

To improve the stability of model evaluation, this paper introduces 5-fold cross validation in the training process. The training set is divided into five mutually exclusive subsets, one of which is selected as the verification set each time, and the other four are used as the training data, and the subsets are rotated after the training is completed. Finally, the average performance index of five experiments is taken as the final evaluation result of the model. This method can effectively reduce the contingency caused by single data division and improve the statistical reliability of experimental results.

To verify the generalisation ability of the model in different dance types and scenes, this study also designed a scene verification experiment: testing in four typical scenes of indoor stage, outdoor festival, dense crowd and drastic lighting changes, and counting the difference of segmentation accuracy. At the same time, the video with different resolutions (720p, 1080p, 4K) and frame rates (24 fps, 30 fps, 60 fps) are tested in groups to analyse the adaptability of the model to the changes of shooting conditions. In the training process, all experiments are carried out under the same hardware and hyperparameter settings to ensure the comparability of the results. The training hardware is a workstation equipped with Nvidia RTX 4090 GPU, the software framework is PyTorch 2.0, the batch size is 8, the number of training rounds is 100, and the learning rate is 1×10^{-4}.

2.3.3 Comparing the algorithm with the baseline model

The validity of the proposed model is verified, and two mainstream image segmentation models are selected as comparison algorithms, and a classic baseline method is set up.

- U-Net (baseline): as a classic framework of medical image segmentation, U-Net has good adaptability to small sample scenes, but in scenes with complex background and changeable movements such as ethnic dance videos, there are problems of blurred boundaries and local missed detection.
- DeepLabv3+: as the mainstream semantic segmentation network, DeepLabV3+ adopts hole convolution and multi-scale context feature fusion, which is suitable for dealing with targets with scale changes.

This study uses its standard version as a comparison. Proposed model (the method in this paper): multi-scale feature fusion and attention mechanism are introduced on the basis of U-Net, and the motion detection prior is combined to improve the segmentation performance in complex scenes.

In the experiment, the above three models are all trained under the same dataset and training strategy, and their performance is evaluated on the same test set to ensure the fairness of comparison. For U-Net and DeepLabv3+, their public implementation and recommended superparameters are used for reproduction, and the input resolution and loss function weight are adjusted appropriately for this dataset. An ablation version without motion detection module is also set up, that is, the original video frame is directly input into the segmentation network to evaluate the improvement of the final performance by the motion area detection module. By comparing the index differences

between the three main models and the ablation version, the contribution of each module can be clearly analysed.

2.3.4 Statistical analysis methods

To ensure the scientificity and credibility of the experimental results, various statistical analysis methods are introduced into the performance evaluation. Firstly, the average and standard deviation of all performance indexes obtained by 50% cross-validation are calculated to evaluate the stability of the model under different training divisions. According to the performance differences among different algorithms (U-Net, DeepLabv3+, and the method in this paper), the paired t-test is used to judge whether they are statistically significant.

- Assumption h_0: There is no significant difference in the average performance index between the two algorithms.
- Alternative hypothesis h_1: There are significant differences between the two algorithms.

After calculating the value of p, if $p < 0.05$, the difference is considered statistically significant. All statistical tests are implemented using SciPy 1.11.

The performance differences of the model in different scenes and dance types were analysed, and the differences between the IoU and Dice mean values in four kinds of scenes (indoor, outdoor, crowded, lighting changes) were compared by using the one-way analysis of variance (ANOVA). If the difference is significant, Tukey HSD back testing is performed to determine the specific source of the difference.

3 Results and analysis

3.1 Experimental results and performance analysis

3.1.1 Comparative analysis of segmentation performance of different algorithms

The overall performance of the dynamic segmentation algorithm proposed in this paper is verified, and the system comparison experiments are carried out with the classic U-Net and the mainstream DeepLabv3+ model on the same dataset. All three algorithms adopt the same training strategy and super parameter settings, and the evaluation indexes include you, Dice, precision, recall and F1-score, which fully reflect the spatial overlap, boundary consistency and classification accuracy of segmentation performance.

Table 4 Comparison of segmentation accuracy of different algorithms

Algorithm	*IoU*	*Dice*	*Precision*	*Recall*	*F1-score*
U-Net	0.82	0.85	0.88	0.83	0.85
DeepLabv3+	0.84	0.87	0.89	0.86	0.87
The method in this paper	0.89	0.91	0.93	0.9	0.91

Table 4 shows that our method is superior to U-Net and DeepLabv3+ in all indicators, and the IoU index is 0.89, which is 7 percentage points higher than U-Net and 5 percentage points higher than DeepLabv3+, indicating that our method has higher overlapping accuracy in spatial segmentation of dancers' foreground and background areas. Dice coefficient is also increased from 0.85 of U-Net and 0.87 of DeepLabv3+ to 0.91, which shows that the algorithm can extract action areas more completely in dance videos with elegant costumes and overlapping characters. In precision and recall, the method in this paper reaches 0.93 and 0.90, respectively, and the F1-score is 0.91, which shows that a good balance is achieved between the accuracy and recall of the positive category (dancer area).

The performance improvement mainly benefits from two aspects:

1 the motion area detection module pre-screened the dancer area in the early stage, which effectively reduced the background interference and improved the quality of the model's attention area

2 multi-scale feature fusion and attention mechanism strengthen the perception ability of dancers' limb edges and clothing details, making the boundary segmentation more complete and less false positives.

By visually comparing the effects of different algorithms, Figure 2 shows the segmentation visualisation results in a typical outdoor festival dance scene.

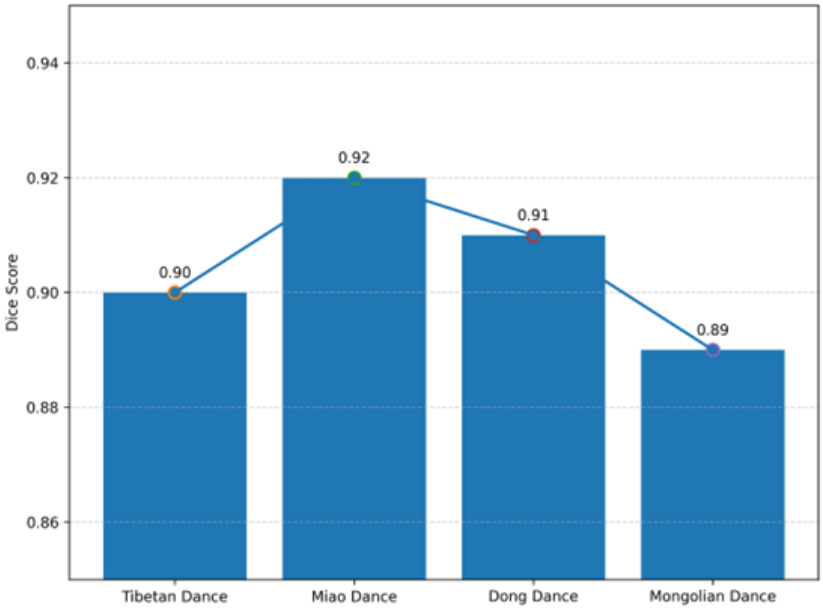
Figure 2 Visualisation of segmentation effect of different algorithms in festival dance video (see online version for colours)

As can be seen from Figure 2, there are obvious phenomena of 'breaking' and 'missing points'. In U-Net at the junction between dancers and background; DeepLabv3+ has improved the problem, and there are still mistakes in clothing edges and uneven lighting areas; the method in this paper maintains good contour integrity and background suppression effect in dancers' costumes, limbs and overlapping action areas, which proves the effectiveness of multi-module combination.

3.1.2 Analysis of segmentation performance under different dance types

There are significant differences in action range, rhythm characteristics and costume form among different ethnic dances, which have a direct impact on the performance of image segmentation algorithm. In this paper, four typical dances (Tibetan pot dance, Miao Lusheng dance, Dong Da Ge dance and Mongolian Andai dance) are selected, and the Dice coefficient on the test set is counted and the comparison chart is drawn.

Figure 3 Comparison of segmentation performance of different dance types (see online version for colours)



As shown in Figure 3, Miao dance (0.92) and Dong dance (0.91) are better than Tibetan dance (0.90) and Mongolian dance (0.89). The reason is that Miao dance and dong dance are usually group dances with stable rhythm, clear outline of costumes and strong regularity of dancers' movements in the picture. However, Tibetan Guozhuang dance rotates frequently, Mongolian Andai dance has a large range, and there are many fast body movements, which are easy to cause motion blur and edge aliasing, which increases the difficulty of segmentation. The method in this paper maintains high accuracy in all four kinds of dances, and the model has good adaptability under different dance action characteristics, but it also reveals that there is room for further optimisation in high-rotation and high-speed action scenes, such as introducing time consistency constraints and time sequence feature modelling.

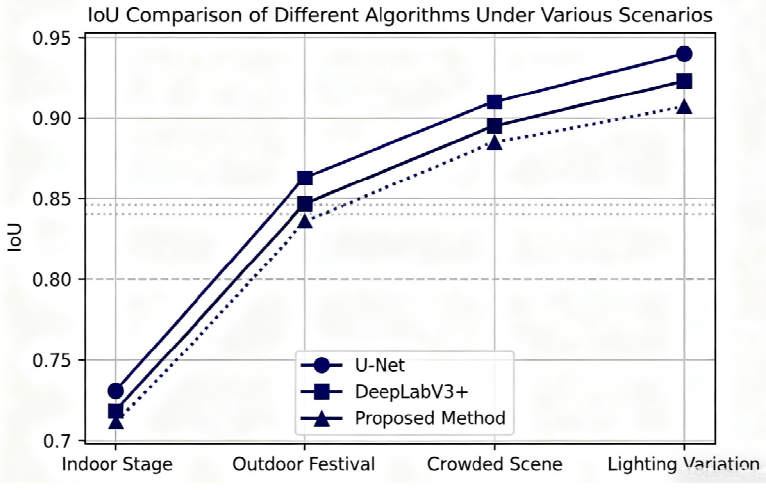
3.1.3 Robustness analysis under different scenes and lighting conditions

National dance performances are usually performed in a variety of scenes, including stage performances, outdoor festivals and crowded folk occasions. The illumination, background complexity and occlusion in different scenes are quite different, which puts forward higher requirements for the robustness of the segmentation algorithm. In this

paper, four typical scenarios are selected, and the IoU performance of U-Net, DeepLabv3+ and the method in this paper are counted respectively.

It can be seen from Figure 4 that the performance of the three algorithms decreases with the increase of the complexity of the scene, but the method in this paper always maintains obvious advantages. In the scene of changing illumination and crowded people, the IoU of U-Net is reduced to 0.72, and DeepLabv3+ is maintained at 0.78, and the method in this paper still maintains a high level of 0.84. This shows that the algorithm in this paper has stronger stability when dealing with complex background, dynamic illumination and occlusion. The main reason is that the pre-motion detection module can effectively filter out most of the background interference areas, while the multi-scale feature fusion mechanism enhances the comprehensive understanding ability of the model to local and global information, so that the algorithm can maintain good segmentation effect in unstructured real shooting environment.

Figure 4 IoU comparison of three algorithms in different scenarios (see online version for colours)

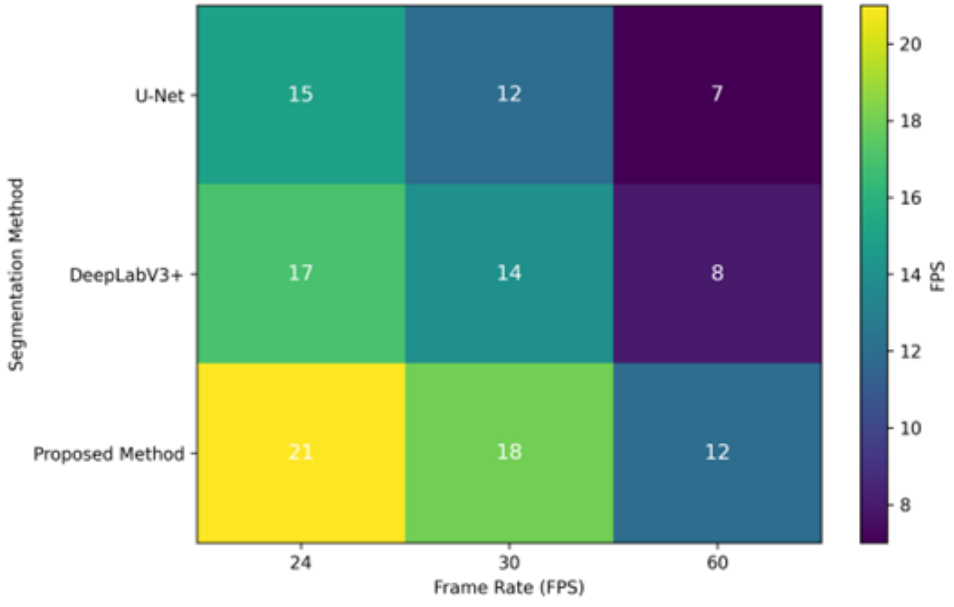


3.1.4 Real-time processing performance analysis under different video frame rates

In the actual digital protection of national dance and teaching scene, the processing speed of the algorithm is equally critical. In this section, the processing speeds of the three algorithms are measured at three common video frame rates: 24 fps, 30 fps and 60 fps.

As shown in Figure 5, the processing speed of this method reaches 21 fps and 18 fps in 24 fps and 30 fps scenes, respectively, which is obviously better than U-Net (15 fps and 12 fps) and DeepLabv3+ (17 fps and 14 fps), and is close to the requirements of real-time processing. Under the high frame rate of 60 fps video, the processing speed of the three algorithms has decreased, but the method in this paper is still 71% higher than that of U-Net and 50% higher than that of DeepLabv3+. This shows that the method in this paper gives consideration to reasoning efficiency while maintaining high accuracy, and provides support for online identification, teaching interaction and digital display of non-legacy dances.

Figure 5 Processing speed (fps) at different frame rates (see online version for colours)



3.2 Discussion

3.2.1 Problems and challenges encountered in the research

This study has achieved good experimental results in the process of dynamic segmentation of non-legacy ethnic dance videos, but it still faces many challenges in research and practical application. The cost of data collection and labelling is high, and ethnic dances often occur in real performance scenes. The illumination, background, crowd and camera angle are not controlled, and the collected videos have problems such as great quality difference, serious motion blur and frequent local occlusion. These factors directly affect the segmentation accuracy, and also lead to a huge workload of data preprocessing and cleaning. Accurate pixel-level labelling requires manual drawing of the dancer's outline frame by frame, and the time cost is extremely high, especially in the case of multi-person group dancing and complex costumes, it is more difficult to label.

Motion blur and non-rigid motion lead to unstable segmentation boundary. Some dances, such as Tibetan pot dance and Mongolian Andai dance, have large-scale rotation and rapid limb movements, which easily lead to obvious blurring between frames, leading to fuzzy response or misjudgement of traditional detection methods based on spatial gradient and optical flow. Even if the depth model is combined, artefacts and jitter will occur at the edges. In addition, dancers' costumes are often wide and elegant, and their edge regions show non-rigid deformation in different frames, which brings extra difficulty to the feature learning of segmented networks.

The robustness of the algorithm is still limited in extreme scenes. For example, in night festival scenes, there are complex lighting changes and large shadows, and the brightness of some dancers' body areas is close to the background, which makes it

difficult for motion detection and segmentation networks to distinguish between real foreground and noise areas. In crowded folk performances, the overlapping of dancers' limbs will also cause confusion in prediction and 'foreground adhesion'.

The balance between real-time and accuracy is also a challenge. Although this method is close to real-time processing at 30 fps, the processing speed is still low in 60 fps or even higher resolution video, which is difficult to meet the needs of large-scale real-time scene acquisition. The improvement of segmentation accuracy will lead to the increase of network structure complexity and computational overhead. How to improve the reasoning efficiency while ensuring the boundary quality is a technical bottleneck that the future application of the algorithm must face. While the proposed model achieved promising segmentation accuracy, several methodological limitations were identified. The system shows reduced stability under extremely low lighting or scenes with heavy costume occlusion, where motion boundaries become indistinct. In highly crowded choreographies, partial overlaps between dancers can still cause misclassification of body regions. Moreover, the reliance on frame-based features may limit performance when temporal consistency is disrupted by rapid transitions. Future research should explore adaptive illumination compensation and spatiotemporal attention mechanisms to enhance robustness. Recognising these limitations not only strengthens the transparency of the study but also provides valuable guidance for refining intelligent segmentation systems in cultural contexts.

3.2.2 Suggestions and improvement directions for future research

In order to solve the above problems, the future research will be improved and expanded from three levels: data, algorithm and application. First, a larger and standardised dataset of non-legacy national dance videos will be established on the data level, covering the real performance environment of multi-ethnic, multi-angle, multi-illumination and multi-scene to enhance the generalisation ability of the model. In order to reduce the labelling cost, semi-automatic or weakly supervised labelling method can be introduced, which combines optical flow, key point detection and automatic contour extraction for initial labelling, and then manual correction can be carried out to improve efficiency. Multi-view synchronous acquisition and 3D reconstruction technology can also be used to generate high-precision dancer contours and provide strong prior information for segmentation.

Time series modelling and cross-frame consistency constraints are introduced in the algorithm level, mainly for single frame segmentation, and there is no explicit modelling of dance movement time series information, which leads to a slight decline in the effect in fast-moving and blurred boundary scenes. In the future, we can combine ConvLSTM, transformer or spatio-temporal attention mechanism, and introduce timing consistency constraints to reduce cross-frame jitter and artefacts on the premise of ensuring real-time. Explore lightweight network structure (such as MobileNet, EfficientNet), model pruning and quantisation technology, reduce reasoning delay under the premise of maintaining accuracy, and meet the processing requirements of high frame rate scenes.

At the application level, the segmentation results are combined with other tasks (such as posture estimation, motion recognition and dance semantic annotation) to build a more perfect digital protection and dissemination platform for national dances. For example, the segmented dancer region can be input into the pose estimation network, which can realise the digital modelling of traditional dance movements; combined with the motion

recognition module, dance motion labels and teaching tips can be automatically generated, which provides technical support for non-legacy dance teaching and interactive display. In the future, we can also consider deploying a lightweight segmentation model on edge devices to meet the real-time application needs in rural areas or festivals.

Beyond the empirical results, this study contributes theoretically to the intersection of digital heritage preservation and intelligent image segmentation. The proposed model demonstrates that computational algorithms can serve as mediators between cultural expression and technological representation. By integrating motion detection and deep learning, the framework not only enhances segmentation performance but also supports the cultural fidelity of dance representation. Theoretically, it suggests that algorithmic design can embed heritage-sensitive principles, ensuring that digitisation processes respect cultural meaning. From a broader perspective, this approach provides a conceptual bridge between computational modelling and cultural semiotics, advancing interdisciplinary dialogue in digital humanities and computer vision.

4 Conclusions

Focusing on the digital protection and inheritance requirements of national dance in intangible cultural heritage, aiming at the problems of traditional video segmentation methods in complex dance scenes, such as insufficient accuracy, poor robustness and weak real-time performance, this paper puts forward an image dynamic segmentation algorithm and a complete technical process for non-legacy national dance videos. By constructing a large-scale dance video dataset with multi-ethnic, multi-scene and multi-illumination conditions, and designing modules such as motion area detection, depth segmentation, multi-scale feature fusion and joint loss optimisation, this study has realised high-precision automatic segmentation of dancers' motion areas in real and complex performance scenes. Compared with mainstream algorithms such as U-Net and DeepLabv3+, the experimental part shows that the proposed method has improved in IoU, Dice, precision, recall and F1-score, especially in complex scenes such as group dance, lighting changes and crowded people. The processing speed of this method under the condition of 30 fps video reaches 18 fps, which is close to the real-time level, which proves its practical application potential in online identification and teaching scene of national dance.

The innovation is embodied in three aspects. First, the motion area detection module combining frame difference and optical flow is introduced to effectively suppress background interference before depth segmentation and enhance the spatial focusing ability of the algorithm. Secondly, the multi-scale feature fusion and attention mechanism are applied to the national dance scene, which significantly improves the ability to capture the edge of clothing and body movements; thirdly, the joint loss function of Dice and cross entropy is used to optimise the sparse foreground region effectively, which makes the segmentation result better than the traditional method in contour integrity and regional consistency.

The future research will be further expanded around three directions. First, on the data level, a more standardised multi-ethnic and multi-angle annotation video database will be established, and semi-automatic annotation and 3D reconstruction technology will be introduced to reduce the annotation cost and improve the annotation accuracy. The

second is to explore time series modelling and lightweight network structure at the algorithm level, and introduce methods such as transformer and ConvLSTM to enhance cross-frame consistency and real-time. Thirdly, at the application level, the segmentation results are combined with gesture recognition, action analysis and digital twin dance model to build a comprehensive digital intangible cultural heritage platform for teaching, inheritance and display, so as to realise the precise protection and innovative communication of national dance. The dynamic segmentation method proposed in this paper not only provides technical support for the automatic processing of complex dance videos, but also opens up a new way for the digital protection and inheritance of national cultural heritage, which has academic value and application prospects.

This research advances both the technological and interdisciplinary understanding of ethnic dance video analysis. It critically demonstrates that integrating motion detection, multi-scale feature fusion, and cultural semantics enhances the authenticity and precision of digital representations. Nevertheless, the study acknowledges certain limitations, such as sensitivity to extreme lighting and the need for larger datasets covering more diverse ethnic traditions. Future research should explore multimodal approaches that combine audio-visual and kinematic data to enrich the interpretive capacity of segmentation systems. Furthermore, cross-disciplinary collaboration among engineers, anthropologists, and cultural practitioners is essential to ensure that digital technologies serve not merely as analytical tools but as platforms for sustaining and revitalising intangible cultural heritage. This research demonstrates that integrating deep learning-based segmentation with heritage-sensitive design principles can enhance both accuracy and cultural authenticity in ethnic dance digitisation. The proposed framework not only achieves precise boundary recognition and motion tracking but also supports broader applications in education, digital archiving, and public engagement. For instance, segmented dance data can serve as training material for cultural education programs or as digital assets for immersive exhibitions and VR-based learning environments. Looking forward, future research may adapt this method to evolving recording technologies – such as high-frame-rate and multi-view capture systems – to capture finer motion details and multidimensional choreography. Ultimately, this approach envisions a dynamic interaction between technology and tradition, ensuring that algorithmic advancement contributes to the living continuity of intangible cultural heritage.

This study presents a novel dynamic segmentation framework that integrates technical innovation with cultural preservation principles. Its novelty lies in bridging the gap between intelligent video analysis and the expressive semantics of ethnic dance. The main contribution is twofold: advancing multi-scale motion segmentation for non-rigid, complex performances, and embedding heritage-sensitive considerations into algorithmic design. Beyond the technical outcomes, this research encourages interdisciplinary collaboration among engineers, anthropologists, and educators. Such cooperation can transform the proposed framework into practical tools for education, digital archiving, and cultural exhibition. In doing so, it contributes not only to computer vision but also to the sustainable transmission of intangible cultural heritage in the digital age.

Declarations

The author declares that he has no conflicts of interest.

References

Abbasimehr, H. and Bahrini, A. (2022) 'An analytical framework based on the recency, frequency, and monetary model and time series clustering techniques for dynamic segmentation', *Expert Systems with Applications*, 15 April, Vol. 192, p.116373.

Davis, S.K. (2020) 'Dancing in the street: impacting at-risk youths' lives through the arts', *Sociological Perspectives*, Vol. 63, No. 3, pp.516–518.

De Souza, E.A.G., Nagano, M.S. and Rolim, G.A. (2022) 'Dynamic programming algorithms and their applications in machine scheduling: a review', *Expert Systems with Applications*, 15 March, Vol. 190, p.116180.

Fan, Q. (2024) 'Research on intangible cultural heritage resource description and knowledge fusion based on linked data', *The Electronic Library*, Vol. 42, No. 4, pp.521–535.

Fan, Q., Sun, C.M. and Zhang, M. (2023) 'Research on the knowledge organization of intangible cultural heritage spatiotemporal data from a digital humanities perspective', *Knowledge Organization*, Vol. 50, No. 8, pp.526–541.

Fan, Q., Tan, G.X., Sun, C.M. and Chen, P.F. (2022) 'Research on knowledge organization of intangible cultural heritage based on metadata', *Information Technology and Libraries*, Vol. 41, No. 2, pp.1–13.

Gao, Y.Y., Zhu, Z.H., Fang, F., Zhang, Y.C. and Meng, M. (2024) 'EEG emotion recognition based on data-driven signal auto-segmentation and feature fusion', *Journal of Affective Disorders*, 15 September, Vol. 361, pp.356–366.

Kaholokula, J.K., Look, M., Mabellos, T., Ahn, H.J., Choi, S.Y., Sinclair, K.A. et al. (2021) 'A cultural dance program improves hypertension control and cardiovascular disease risk in native Hawaiians: a randomized controlled trial', *Annals of Behavioral Medicine*, Vol. 55, No. 10, pp.1005–1018.

Li, Y. and Gang, J.L. (2021) 'Development of art and culture creative industry using FPGA and dynamic image sampling', *Wireless Communications and Mobile Computing*, Vol. 2021, p.6639045.

Luo, J.M., Hu, Z.W. and Leong, A.M.W. (2025) 'Exploring the experience attributes of intangible cultural heritage through big data analytics', *Journal of Vacation Marketing*, p.13567667251323644.

Ma, T., Dang, Z.R., Yang, Y.Z., Yang, J.Y. and Li, J.H. (2024) 'Dental panoramic x-ray image segmentation for multi-feature coordinate position learning', *Digital Health*, Vol. 10, p.20552076241277154.

Mao, Q., Mastnak, W. and Guan, R.Y. (2025) 'Chinese ethnic dance therapy: cultural anthropology and health science perspectives on Tujia ethnic dances', *Frontiers in Psychology*, Vol. 16, p.1561150.

Munusamy, S. and Murugesan, P. (2020) 'Modified dynamic fuzzy c-means clustering algorithm – application in dynamic customer segmentation', *Applied Intelligence*, Vol. 50, No. 6, pp.1922–1942.

Sjobeck, G.R., Boker, S.M., Scheidt, C.E. and Tschacher, W. (2024) 'The pairwise approximate spatiotemporal symmetry algorithm: a method for segmenting time series pairs', *Psychol. Methods*, Vol. 29, No. 3, pp.435–456.

Song, R., Qin, W.E., Shi, W. and Xue, X.J. (2023) 'Optimizing freight vehicle routing in dynamic time-varying networks with carbon dioxide emission trajectory analysis', *Sustainability*, Vol. 15, No. 21, p.15504.

Su, L.J., Wang, W.Y. and Xu, X.B. (2023) 'Identifying latent group structures in spatial dynamic panels', *Journal of Econometrics*, Vol. 235, No. 2, pp.1955–1980.

Wang, J.C., Chen, M., Zhang, H.Y. and Ye, F. (2023) 'Intangible cultural heritage in the Yangtze river basin: its spatial distribution characteristics and influencing factors', *Sustainability*, Vol. 15, No. 10, p.7960.

Wang, Y. (2025) 'Promoting student mental health through school music and dance education in ethnic cultural inheritance', *The International Journal of Psychiatry in Medicine*, Vol. 60, No. s4, pp.167s–169s.

Wu, L.X., Yang, G.L. and Chen, X.W. (2024) 'Spatial distribution characteristics and influencing factors of intangible cultural heritage in the Yunnan, Guangxi, and Guizhou rocky desertification are', , Vol. 16, No. 11, p.4722.

Zhao, M. (2022) 'The influence of ethnic dance education in colleges and universities on alleviating college students' mental anxiety', *Psychiatria Danubina*, Vol. 34, No. Suppl. 2, pp.S258–S259.