

Smart Innovation, Systems and Technologies 500

Roumiana Kountcheva
Srikanta Patnaik
Yingkai Liu *Editors*



Advances in 3D Technologies

Proceedings of 7th International
Conference on 3D Imaging
Technologies 2025, Volume 2



 Springer

Smart Innovation, Systems and Technologies

500

Series Editors

Robert J. Howlett, *KES International, Shoreham-by-Sea, UK*

Lakhmi C. Jain, *KES International, Shoreham-by-Sea, UK*

The Smart Innovation, Systems and Technologies book series encompasses the topics of knowledge, intelligence, innovation and sustainability. The aim of the series is to make available a platform for the publication of books on all aspects of single and multi-disciplinary research on these themes in order to make the latest results available in a readily-accessible form. Volumes on interdisciplinary research combining two or more of these areas is particularly sought.

The series covers systems and paradigms that employ knowledge and intelligence in a broad sense. Its scope is systems having embedded knowledge and intelligence, which may be applied to the solution of world problems in industry, the environment and the community. It also focusses on the knowledge-transfer methodologies and innovation strategies employed to make this happen effectively. The combination of intelligent systems tools and a broad range of applications introduces a need for a synergy of disciplines from science, technology, business and the humanities. The series will include conference proceedings, edited collections, monographs, handbooks, reference books, and other relevant types of book in areas of science and technology where smart systems and technologies can offer innovative solutions.

High quality content is an essential feature for all book proposals accepted for the series. It is expected that editors of all accepted volumes will ensure that contributions are subjected to an appropriate level of reviewing process and adhere to KES quality principles.

Indexed by SCOPUS, EI Compendex, INSPEC, WTI Frankfurt eG, zbMATH, Japanese Science and Technology Agency (JST), SCImago, DBLP.

All books published in the series are submitted for consideration in Web of Science.

Roumiana Kountcheva · Srikanta Patnaik ·
Yingkai Liu
Editors

Advances in 3D Technologies

Proceedings of 7th International Conference on 3D
Imaging Technologies 2025, Volume 2

Editors

Roumiana Kountcheva
TK Engineering
Sofia, Bulgaria

Srikanta Patnaik
Institute of Management and Technology
Bhubaneswar, Odisha, India

Yingkai Liu
Yunnan Normal University
Kunming, Yunnan, China

ISSN 2190-3018

ISSN 2190-3026 (electronic)

Smart Innovation, Systems and Technologies

ISBN 978-3-032-25612-6

ISBN 978-3-032-25613-3 (eBook)

<https://doi.org/10.1007/978-3-032-25613-3>

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2026

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface

The 7th International Conference on 3D Imaging Technologies - Digital Signal Processing and Deep Learning (3DIT-DSP&DL) was successfully held in Sanya, China, during November 21–23, 2025. The conference topics cover all aspects of theories and applications in image processing and multidimensional signal processing and analysis, in a wide variety of related applications, based on Deep-learning, AI technologies, and virtual reality.

The 3DIT-DSP&DL conference became already a traditional and significant event in the related scientific areas and attracted authors and keynote speakers from China, Japan, Belarus, Singapore, USA, UK, Canada, Jordan, Ukraine and Bulgaria. After serious analysis and evaluation, 66 papers were selected for publication in the two volumes of the conference proceedings.

This book is the second volume of the conference proceedings, which contains 29 chapters, arranged in three groups in accordance with their scientific focus:

- I. 3D technologies in art and tourism (8 chapters);
- II. Key technology solutions (13 chapters);
- III. Intelligent education and training systems (8 chapters).

We wish to express our sincere appreciation to all participants and the technical program committee for their efforts to review and evaluate all submissions, which is vital for the successful conduct of the conference 3DIT-DSP&DL and also to the organizer's team, who had dedicated their time and efforts in planning, promoting, organizing, and supporting the conference. Our special appreciation is extended to the keynote and invited speakers Prof. Mohammad S. Obaidat (University of Jordan, Jordan), Prof. Andrii Bieliatynskyi (North Minzu University, China), Prof. Kun Zhang (Hainan Normal University, China) and S.R. Roumiana Kountcheva (TK Engineering, Bulgaria), who made valuable and beneficial speeches for the conference audience. We also thank the Springer publishing team, which made this book possible.

January 2026

Roumiana Kountcheva
Srikanta Patnaik
Yingkai Liu

Contents

3D Technologies in Art and Tourism

Research on Interactive Display Design of Product Three-dimensional Digital Prototype Under the Background of Cultural and Tourism Integration	3
<i>Yuhang Wang and Wei Zhang</i>	
Research on Human-Computer Interactive Communication of Intangible Cultural Heritage Three-Dimensional Digital Technology Driven by Artificial Intelligence in the Context of Cultural and Tourism Integration	15
<i>Wei Zhang, Yuhang Wang, Jinyou Wang, and Xiaotong Wang</i>	
Research on the Integrated Development of Public Art in Urban Public Spaces Empowered by Digital Technology	27
<i>Chao Yuan and Hua Zhu</i>	
Research on the Application of Modern Exhibition Hall Visual Design Integrating VR, AR and AI Technologies	38
<i>Jiafeng Wu, Zishan Zeng, and Yixuan Wang</i>	
Digital Preservation and Display of Intangible Cultural Heritage Techniques Based on 3D Reconstruction and Virtual Reality	49
<i>Yanan Cai, Yujia Liang, and Qingyin Zhang</i>	
Research on the Application of 3D Modeling Software CINEMA4D + Octane Render in Product Creative Design Courses	61
<i>Yuhang Wang, Wei Zhang, Wenya Zhang, and Haoyun WangXin</i>	
Application of Computer System Virtualization Technology in IT Architecture Optimization for Cultural and Tourism Enterprises	70
<i>Yanan Cai, Yudi Wang, and Weize He</i>	
Research on the Innovation of Zhuanghe Paper-Cutting Intangible Cultural Heritage Enabled by Generative AI Based on SEM Model	82
<i>Wei Zhang, Qingying Zhou, Yajun Chen, and Yumeng Cuan</i>	

Key Technology Solutions

EPKG: A Energy Policy Knowledge Graph Construction Approach	95
<i>Xuepeng Mou, Julong Chen, Bin Wang, Wei Wang, Chen Luo, Zhen Li, Yongqing Zhu, Shiping Yang, and Rui Hua</i>	
CrossLiteCrackNet: A Lightweight Deformable Attention Network for Real-Time Pavement Distress Segmentation	110
<i>Rongyan Guo, Peizhi Tian, and Yafei Wang</i>	
Quantitative Analysis of Elite Coaching Effects in Olympic Sports: A Hybrid CUSUM-OLS Approach	126
<i>Tie Yan, Ruiqi Wang, Yunxuan Gao, and Shiqi Wang</i>	
Ludonarrative Harmony: Designing an AI-Assisted Framework for Level Design Decision Support	138
<i>Dongmeng Zhang, Luying Yang, and Yufei Xue</i>	
Standardized Evaluation System of Traditional Martial Arts Actions Based on Deep Learning and Visual Processing Algorithms	153
<i>Shaojie Li</i>	
Research on Digital Reconstruction and Optimization of Business Governance Processes in Ultra-high Voltage Grids	165
<i>Xuefei Gao, Jianlin Wang, Liuqi Zhao, Yuheng Zhang, Zhihui Chen, and Xing Wen</i>	
Performance Evaluation of Exploration-Exploitation Strategies in Genre-Level Movie Recommendation: A Multi-armed Bandit Perspective	176
<i>Yisong Li</i>	
Short-Term Traffic Flow Forecasting Model Incorporating Spatio-Temporal Graph Convolution	188
<i>Hao Wei</i>	
Comparative Analysis of Exploration Strategies in Stationary Multi-armed Bandit Problems: An Empirical Study of ϵ -Greedy, UCB1, and Thompson Sampling	200
<i>Mingyang Xu</i>	
Intelligent Quality Monitoring and Autonomous Decision Support System for Tobacco Production Workshops	212
<i>Xiaofei Chen, Guoyang Lu, Liang Ping, and Changyong Yang</i>	

Research on Key Technologies of 3D Rapid Reconstruction Based on Multi-modal Deep Learning	225
<i>Ling Peng, Shengli Yang, and Tianyi Liu</i>	
Application of Intelligent Image Recognition System Based on Deep Learning in Upstream Oil and Gas	236
<i>Jun Li, Xiangyang Zhang, Zhanbao Fu, and Ying Qi</i>	
High-Voltage Switchgear Lifecycle Intelligent Operation and Maintenance System Based on Digital Twin	245
<i>Meng Jin, Jiangtao Xia, Linqi Xie, Teng Li, and Shangduo Zhang</i>	
Education and Training Systems	
Intelligent Rehabilitation-Action Evaluation System via Enhanced DTW and Deep Learning	255
<i>Yizhou Han, Yijia Xu, Haonan Lu, and Yongjun Gan</i>	
Research on Emergency Rescue Technology Based on Digital Twin	275
<i>Weilun Guo, Weimin Li, Xinxing Mu, and Liang Ye</i>	
Application Scenario Design of VR/AR Technology in Early Infant Education	288
<i>Ting Zhang</i>	
Multimodal Intelligent Evaluation System for Spoken English Expression Based on Audio and Text	301
<i>Xiao-xiao Duan and Tong Zhou</i>	
Design of Elderly Care System Based on Deep Learning and Expression Recognition Technology of Artificial Intelligence	313
<i>Feng-Hsiung Hou</i>	
Internal Deformation Measurement of 4D-Printed Structures Using Computed Tomography (CT) and Digital Volume Correlation (DVC)	325
<i>Zijia Ma, Zhongwei Li, Kai Zhong, and Xin Jiang</i>	
Construction and Research of a Multi-dimensional Real-Time Feedback System for English Writing Based on Transformer	336
<i>Xiao-xiao Duan, Ting-ting Duan, and Ping Duan</i>	

Research on the Design of Intelligent Evaluation Module of Online
Learning Platform Based on Deep Learning 349
 Xuefeng Xin

Author Index 361

About the Editors

Prof. Roumiana Kountcheva got her M.Sc. and Ph.D. at the Technical University of Sofia, Bulgaria, and in 1992, she got the title senior researcher. At present, she is the vice president of TK Engineering, Sofia. She had postgraduate trainings in Fujitsu and Fanuc, Japan. Her main scientific interests are in image processing, image compression, digital watermarking, pattern recognition, image tensor representation, neural networks, CNC, and programmable controllers. She has more than 220 publications and 5 international patents (EU and USA). R. Kountcheva was the plenary speaker at 28 international scientific conferences and scientific events. She is a member of international organizations: Bulgarian Association for Pattern Recognition, International Research Institute for Economics and Management (IRIEM), the Institute of Data Science and Artificial Intelligence (IDSAI), and is a honorary member of the Honorable Editorial Board of the nonprofit peer reviewed open access IJBST Journal Group.

Prof. Srikanta Patnaik is presently working as the director of Interscience Institute of Management & Technology, Bhubaneswar, India. He has received his Ph.D. (Engineering) on Computational Intelligence from Jadavpur University, India, in 1999. He has supervised 32 Ph.D. theses and more than 60 Master theses in the area of Computational Intelligence, Soft Computing Applications, and Re-Engineering. He is the author of 3 text books and more than 100 edited volumes and few invited book chapters, published by leading international publisher like IEEE, ACM, Elsevier, Springer-Verlag, Kluwer Academic, SPIE, IOS Press, and others. Dr. Patnaik is the editor-in-chief of International Journal of Information and Communication Technology and International Journal of Computational Vision and Robotics published by Inderscience Publishing House, England, and also the editor-in-chief of Book Series on “Modeling and Optimization in Science and Technology” published by Springer, Germany.

Dr. Yingkai Liu is the dean of the Institute of Physics and Electronic Information, Yunnan Normal University. He obtained his MS degree in condensed matter physics from Yunnan University in 1998 and Ph.D. in condensed matter physics in 2002 from Nanjing University. He then did his postdoctoral research in the Center of Super-Diamond and Advanced Films (COSDAF) at City University of Hong Kong before joining Yunnan Normal University. He visited City University of Hong Kong from September to June 2019 and Rice University from February to December 2018. His research interests include nanomaterial synthesis & characterization, and Nano device fabrication for photodetectors, gas sensors, environment and biomedical applications. Up to now, he has published 78 articles in the world-famous journals such as Advanced Materials, Advanced Optical Materials, Sensors and Actuators B, Applied Physics Letters, and Nanomaterials. His papers were cited 1800 times by other people.

Education and Training Systems



Intelligent Rehabilitation-Action Evaluation System via Enhanced DTW and Deep Learning

Yizhou Han¹, Yijia Xu², Haonan Lu³, and Yongjun Gan⁴(✉)

¹ Marshall McLuhan Catholic Secondary School, Toronto, ON M5N 3B1, Canada

² Western Academy of Beijing, Beijing 100083, China

³ RDF International School, Shenzhen 518120, China

⁴ School of Mechanical and Electrical Engineering, Beijing Information Science and Technology University, Beijing 100192, China

2024020137@bistu.edu.cn

Abstract. Focusing on the problems of quantitative difficulties, weak model interpretability and insufficient fine-grained anomaly location in rehabilitation action evaluation, this paper proposes an intelligent evaluation framework that combines improved dynamic time warping and deep representation learning. The system uses monocular RGB video as input, relies on a lightweight pose estimation module to obtain the skeletal sequence, and constructs multi-order kinematic features to characterize the spatial morphology and temporal evolution of the action. On this basis, a dynamic time warping strategy incorporating weighted spatio-temporal metrics and topological constraints is introduced to robustly align and locate action rhythm differences and local anomalies. Experimental results demonstrate that the proposed method outperforms multiple baselines on the public dataset in terms of overall similarity, path regularity and abnormal segment overlap rate. Moreover, it can provide actionable correction suggestions for patients via a semantic feedback mechanism. Ablation correlation analysis further verifies the effectiveness of the synergy of features, constraints and metrics.

Keywords: Rehabilitation action evaluation · dynamic time warping · spatio-temporal feature fusion · unmarked attitude estimation · semantic feedback

1 Introduction

With the aging of the global population, the incidence of neurological diseases, particularly stroke, continues to rise. Studies indicate that approximately 75% to 80% of stroke survivors suffer from upper limb motor deficits, necessitating long-term rehabilitation to reconstruct neuromuscular control pathways [1]. However, traditional rehabilitation relies heavily on manual observation and Fugl-Meyer scale scoring by therapists, a process often criticized for its strong subjectivity and low resolution [2]. While wearable Inertial Measurement Units (IMUs) can provide continuous kinematic data, they are prone to signal drift and require cumbersome setup procedures, limiting their usability

in home settings [3]. Similarly, RGB-D sensors offer a contactless alternative but remain constrained by higher hardware costs and depth noise [4, 5]. Consequently, monocular RGB cameras have emerged as a cost-effective solution for home deployment. However, achieving high-precision, real-time, and clinically interpretable motion assessment from 2D video remains a significant technical challenge [6, 7].

Existing research primarily follows two technical pathways. The first involves template matching methods based on Dynamic Time Warping (DTW) and its variants, which achieve action alignment and scoring by calculating the minimum cumulative distance between a test sequence and a standard template [8]. However, classical DTW relies solely on frame-by-frame Euclidean distance, ignoring joint topology and kinematic context. This often results in alignment paths that deviate from the physiologically feasible region, particularly when facing rhythm changes, local occlusions, or motion compensation [9].

The second pathway utilizes end-to-end deep learning frameworks, such as 3D Convolutional Networks [10], Spatio-Temporal Graph Convolutional Networks (ST-GCN) [11], or Transformer alignment networks [12]. While capable of learning complex spatio-temporal representations from large-scale data, these models frequently suffer from attention drift in clinical environments due to training-test distribution shifts. Furthermore, their output often lacks the interpretability required to provide specific correction suggestions for patients [13]. Although attempts at multi-modal fusion (combining RGB and IMU data [14, 15]) have improved robustness against occlusion, they significantly increase system complexity and synchronization difficulties. Therefore, balancing fine-grained anomaly localization, interpretability, and real-time performance using monocular RGB data remains a critical challenge in the field of rehabilitation action evaluation [16].

To address these challenges, this paper proposes an intelligent rehabilitation action evaluation system that integrates improved DTW with deep representation learning. The main contributions are as follows:

- (1) **Design of Multi-order Kinematic Spatio-temporal Coupling Features:** We embed the static pose, local velocity, and acceleration of joints into a unified representation space to comprehensively describe both motion form and dynamic evolution.
- (2) **Proposal of the STCF-DTW Algorithm:** This algorithm utilizes weighted spatio-temporal Euclidean distance as the inter-frame metric and introduces dual constraints—joint topology and local slope—to significantly enhance alignment accuracy and anomaly localization capabilities.
- (3) **Construction of a Semantic Feedback Mechanism:** We establish a closed-loop “score-location-explanation” mechanism that automatically generates semantic correction prompts, supporting patients in completing high-frequency, accurate, and actionable self-training in unsupervised environments.

The rest of the paper is organized as follows: Sect. 2 details the proposed method, including attitude estimation, feature extraction, and STCF-DTW algorithm design; Section 3 introduces the experimental settings, evaluation indicators, comparison and ablation experiments, and shows the clinical usability evaluation; Section 4 summarizes the entire paper and outlines future work.

2 Method

The intelligent evaluation system proposed in this paper adopts a three-stage cascade architecture (Fig. 1) to systematically achieve high-fidelity pose estimation, multi-stage spatio-temporal feature representation, and dynamic alignment. Using monocular RGB video as input, the framework extracts human kinematic primitives via a lightweight markerless perception module, subsequently fusing static joint poses with dynamic evolution features. Finally, incorporating physiological constraints, the system realizes fine-grained anomaly localization and semantic feedback. The following sections provide a detailed examination of the technical implementation and innovative design of each module.

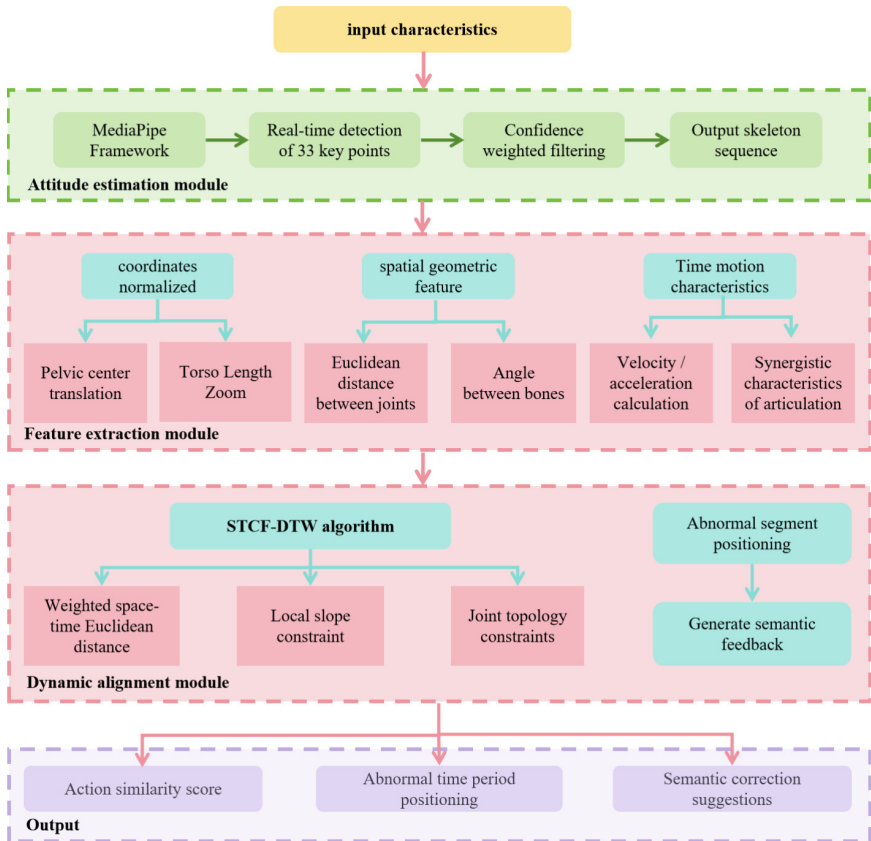


Fig. 1. Model architecture diagram

2.1 High-Fidelity Unmarked Pose Estimation Based on MediaPipe

The system uses Google's MediaPipe framework as the basic pose estimation engine. It captures the patient's rehabilitation actions in real-time via a monocular RGB camera and outputs a sequence of 2D skeleton coordinates for 33 human-body key points. The core advantage of MediaPipe is its lightweight BlazePose model and high inference efficiency ($\geq 30\text{FPS}$), meeting clinical real-time evaluation needs. Additionally, no physical markers are required on the patient's body, greatly reducing operational complexity and equipment costs. The technical process can be summarized as follows: input RGB video frame, human detection based on lightweight CNN, key point positioning based on heat map regression, output normalized two-dimensional coordinate (x_i, y_i) and confidence $c_i \in [0, 1]$ ($i = 0, 2, \dots, 32$ represents joint point index).

The mathematical essence of pose estimation can be expressed as: given the input image I_t at the time step t , model M predicts the joint position:

$$P_t = \{p_t^i | p_t^i = (x_t^i, y_t^i), i = 0, \dots, 32\} \quad (1)$$

The key point confidence c_t^i is obtained by normalizing the maximum response value of the predicted heat map H_i by the Sigmoid function:

$$c_t^i = \sigma(\max_{(u,v)} H_i(u, v)) \quad (2)$$

This confidence quantifies the visibility and positioning reliability of the joint points in the current frame, and provides an important basis for subsequent anomaly detection.

Aiming at the common problems of illumination change and partial occlusion (such as clothing, assistive devices) in rehabilitation scenes, the system introduces a confidence weighted filtering mechanism. The smooth coordinate of the joint point i in the time window τ is defined as:

$$\tilde{p}_t^i = \frac{\sum_{\tau=t-k}^{t+k} w_\tau \cdot c_\tau^i \cdot p_\tau^i}{\sum_{\tau=t-k}^{t+k} w_\tau \cdot c_\tau^i} \quad (3)$$

Where, p_τ^i represents the raw predicted 2D coordinate of joint i at frame τ , and c_τ^i denotes the corresponding confidence score (indicating the reliability of the key point detection). Additionally, w_τ is the Gaussian temporal weight, and k is the window radius⁴. This mechanism effectively suppresses coordinate drift caused by low-confidence frames⁵. The final output sequence $\mathcal{S} = \{\tilde{P}_1, \tilde{P}_2, \dots, \tilde{P}_T\}$ serves as the input for the subsequent module, where T is the total number of action frames⁶. The specific smoothing effects are illustrated in Fig. 2, and the mapping between joint indices and body parts is provided in Table 1.

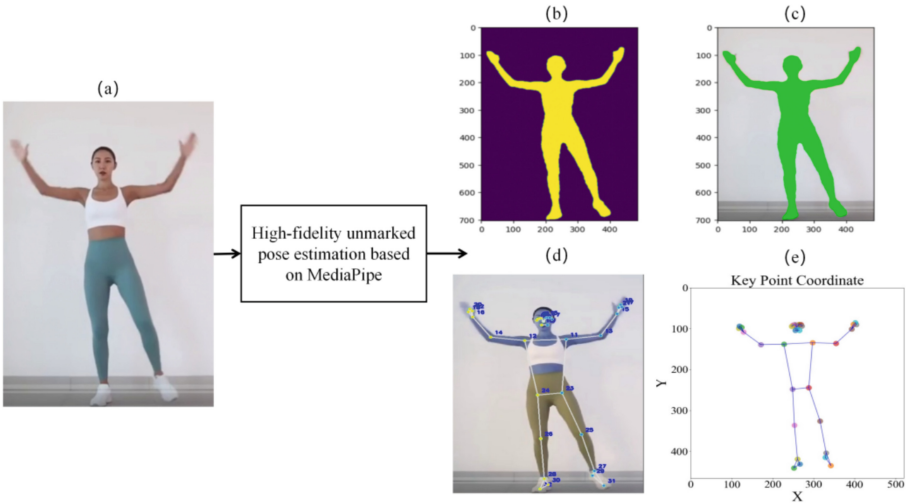


Fig. 2. High-fidelity unmarked pose estimation detection effect map based on MediaPipe

Figure 2 illustrates the markerless pose estimation pipeline. Specifically, Fig. 2(b) and (c) demonstrate background noise suppression achieved via foreground matting, while Fig. 2(d) visualizes the localization of 33 key points and the reconstruction of the skeletal topology. Subsequently, Fig. 2(e) depicts the mapping of these key points into a 2D coordinate system to form a temporally consistent sequence. Additionally, Table 1 lists the key point indices, establishing a unified semantic reference for the subsequent “multi-order kinematic spatio-temporal coupling feature” module.

Table 1. Key points comparison table of high-fidelity unmarked pose estimation algorithm based on MediaPipe

Serial number	Part	Serial number	Part
0	Nose	17	Left pinky
1	Left eye inner	18	Right pinky
2	Left eye	19	Left index
3	Left eye outer	20	Right index
4	Right eye inner	21	Left thumb
5	Right eye	22	Right thumb
6	Right eye outer	23	Left hip
7	Left ear	24	Right hip
8	Right ear	25	Left knee
9	Mouth left	26	Right knee
10	Mouth right	27	Left ankle

(continued)

Table 1. (continued)

Serial number	Part	Serial number	Part
11	Left shoulder	28	Right ankle
12	Right shoulder	29	Left heel
13	Left elbow	30	Right heel
14	Right elbow	31	Left foot index
15	Left wrist	32	Right foot index
16	Right wrist		

2.2 Multi-order Kinematic Spatio-Temporal Coupling Feature Extraction

The core challenge in rehabilitation action assessment lies in accurately quantifying the spatio-temporal deviations between a patient's motion sequence and a standard reference template. Traditional methods, often relying on isolated features, frequently fail to capture complex spatio-temporal coupling dynamics. Furthermore, variations in subject-to-camera distance and inconsistencies in pixel scaling necessitate rigorous coordinate normalization. To address these limitations, we propose a Multi-order Kinematic Spatio-temporal Coupling Feature representation method. This approach integrates static joint pose, velocity, and acceleration to explicitly model both spatial topology and temporal evolution, thereby constructing a robust action descriptor. It effectively mitigates the impact of rhythmic variations, occlusions, and amplitude discrepancies, providing a solid foundation for action classification and anomaly detection. The feature extraction process is organized into three distinct stages.

2.2.1 Coordinates Normalized

In visual motion capture, the patient-camera relative distance dynamically changes, and varying capture-device resolutions lead to inconsistent pixel-coordinate-system scales. To eliminate feature -extraction interference from body-shape differences, perspective changes, and initial-pose deviations, the raw skeleton-coordinate sequence requires strict normalization. Given the input sequence $\mathcal{S} = \{\tilde{\mathbf{P}}_t \in \mathbb{R}^{33 \times 2}\}_{t=1}^T$, where $\tilde{\mathbf{P}}_t = [\tilde{p}_t^0, \tilde{p}_t^1, \dots, \tilde{p}_t^{32}]^\top$ represents the two-dimensional coordinates of 33 key points in the t frame, $\tilde{p}_t^i = (x_t^i, y_t^i)$. In the normalization process, the center of the pelvis is first translated: the center of the pelvis (the midpoint of the left and right hip joints, index $m i = 23, 24$) is selected as the dynamic reference origin, the $\lambda_t = \frac{1}{2}(\tilde{p}_t^{23} + \tilde{p}_t^{24})$ is calculated, and all the joint points in each frame are translated $l_t^i = \tilde{p}_t^i - \lambda_t$. For the scale variation problem, the trunk length (the distance from the left shoulder $i = 11$ to the center of the pelvis $d_t^{\text{torso}} = \|l_t^{11}\|_2$) is used as the reference scale factor, and all the translated coordinates are scaled $\hat{p}_t^{(i)} = l_t^i / d_t^{\text{torso}}$. The normalized coordinate sequence $\hat{\mathcal{S}} = \{\hat{\mathbf{P}}_t\}_{t=1}^T$

has scale invariance and translation invariance, and its mathematical representation is:

$$\hat{P}_t = \frac{1}{d_t^{\text{torso}}} \left(\tilde{P}_t - \begin{bmatrix} \lambda_t^\top \\ \vdots \\ \lambda_t^\top \end{bmatrix}_{33 \times 1} \right) \quad (4)$$

Where, $\tilde{P}_t$ is the key point coordinate matrix of t frame; $\tilde{p}_t^i$ is the two-dimensional coordinate of the i joint point of the t frame; λ_t is the pelvic center coordinate of the t frame; d_t^{torso} is the trunk reference length of the t frame; $\hat{P}_t$ normalized coordinate matrix. The effect of coordinate normalization is shown in Fig. 3.

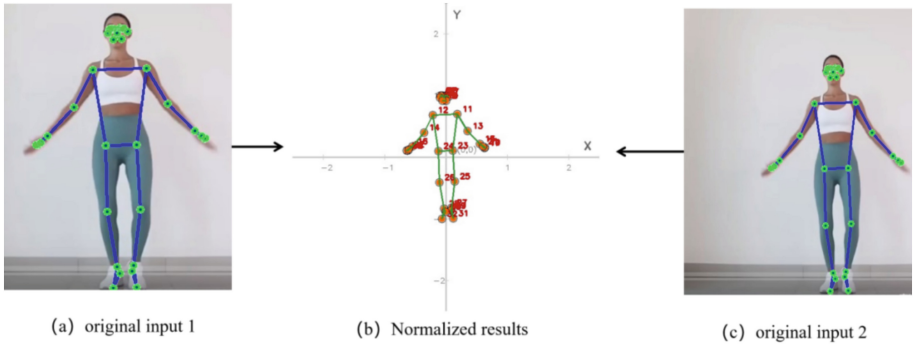


Fig. 3. Coordinate normalization effect

2.2.2 Spatial Geometric Feature Extraction

Based on the normalized joint coordinates, this paper defines two types of core space features: Euclidean distance between joints and angle between bones. These features significantly enhance the ability to describe the action form by explicitly modeling human anatomical constraints, and their physical meaning is clear, providing a direct basis for subsequent abnormal action positioning. The Euclidean distance between the joints represents the degree of spatial separation of specific anatomical point pairs, which is defined as the Euclidean distance between the joint point and the n in the t frame:

$$d_t^{(m,n)} = \|\hat{p}_t^m - \hat{p}_t^n\|_2 \quad (5)$$

For example, the distance $d_t^{(12,16)}$ represents the spatial distance between the right shoulder joint (index 12) and the right wrist joint (index 16) (corresponding to L_1 in Fig. 4), which directly reflects the extension state of the upper arm. The angle between the bones uses the cosine theorem to calculate the joint angle composed of three adjacent joint points to quantify the degree of limb bending. Let the joint points a , b and c form a continuous link in anatomy (b is the vertex of the joint), define the vector $\vec{ba} = \hat{p}_t^a - \hat{p}_t^b$

and $\vec{bc} = \hat{p}_t^c - \hat{p}_t^b$, then the angle $\theta_t^{(b)}$ at the vertex b is calculated as follows:

$$\theta_t^{(b)} = \cos^{-1} \left(\frac{\langle \vec{ba}, \vec{bc} \rangle}{\|\vec{ba}\|_2 \cdot \|\vec{bc}\|_2} \right) \quad (6)$$

Taking the elbow flexion angle as an example (Θ_2 in Fig. 4), it consists of shoulder node (index 12), elbow node (index 14) and wrist node (index 16), $\theta_t^{(14)} = \cos^{-1} \left(\frac{\langle \hat{p}_t^{10} - \hat{p}_t^{12}, \hat{p}_t^{14} - \hat{p}_t^{12} \rangle}{\|\hat{p}_t^{10} - \hat{p}_t^{12}\|_2 \cdot \|\hat{p}_t^{14} - \hat{p}_t^{12}\|_2} \right)$, which directly reflects the flexion and extension range of the elbow joint. In this paper, a total of 8 sets of key distance features and 8 sets of key angle features are defined, covering the main functional links such as upper limbs (shoulder-elbow, elbow-wrist), lower limbs (hip-knee, knee-ankle), and trunk (shoulder-hip). The specific definitions are shown in Table 2. All spatial geometric features are integrated into feature vectors according to frames:

$$\mathbf{f}_t^{\text{space}} = \left[d_t^{(1,2)}, \dots, d_t^{(N_d)}, \theta_t^{(1)}, \dots, \theta_t^{(N_\theta)} \right]^\top \in \mathbb{R}^{N_s}, \quad N_s = 8 + 8 \quad (7)$$

The feature set explicitly encodes the spatial topological rules of human pose through distance constraints and angle constraints. Compared with the original coordinates, it has stronger robustness to individual body shape differences-the distance feature eliminates the scale effect by normalizing the trunk length ratio, and the angle feature has inherent scale invariance. More importantly, its physical meaning has a clear clinical explanation: the therapist can directly identify the lack of limb extension according to the deviation of $d_t^{(m,n)}$, or determine the limitation of joint activity according to the abnormality of $\theta_t^{(b)}$, so as to lay a technical foundation for generating specific correction prompts such as “increasing the bending amplitude of elbow joint”.

Table 2. Spatial geometric features

Normalized results	Key index	Angular characteristics	Key index
L_1	$d_t^{(11,15)}$	Θ_1	$\angle(12 - 14 - 16)$
L_2	$d_t^{(12,16)}$	Θ_2	$\angle(11 - 13 - 15)$
L_3	$d_t^{(13,23)}$	Θ_3	$\angle(13 - 11 - 23)$
L_4	$d_t^{(14,24)}$	Θ_4	$\angle(14 - 12 - 24)$
L_5	$d_t^{(23,27)}$	Θ_5	$\angle(23 - 25 - 27)$
L_6	$d_t^{(24,28)}$	Θ_6	$\angle(24 - 26 - 28)$
L_7	$d_t^{(11,23)}$	Θ_7	$\angle(23 - 24 - 26)$
L_8	$d_t^{(12,24)}$	Θ_8	$\angle(24 - 23 - 25)$

Figure 4 shows the geometric sketches of distance $L_1 = d_t^{(12,16)}$ and angle $\Theta_1 = \theta_t^{(14)}$ as well as $L_5 = d_t^{(23,27)}$ and $\Theta_5 = \theta_t^{(25)}$.

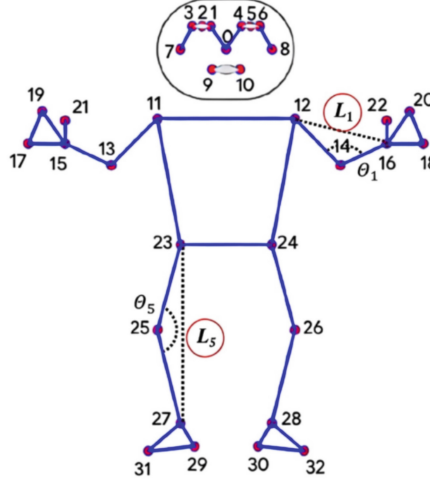


Fig. 4. Schematic diagram of geometric features

2.2.3 Time Motion Feature Extraction

In order to fully describe the dynamic evolution characteristics of rehabilitation actions, it is necessary to further explore the kinematic model in the time series dimension based on the spatial geometric features. The continuity and inertia effect of human motion determine that joint velocity and acceleration are the key indicators to quantify the fluency and coordination of movement. Aiming at the problem that the static features of a single frame cannot reflect the motion trend, and the global statistical features will lose the instantaneous dynamic details, the system introduces a sliding window mechanism (as shown in Fig. 5) to dynamically calculate the local motion features in a time series window with a length of N . The mechanism realizes the fine modeling of the dynamic characteristics of the action by storing the normalized coordinate sequence of continuous N frames. The window slides in real time with the video stream. When the new frame data enters the head, the old data at the tail is automatically eliminated to ensure the timeliness of feature calculation.

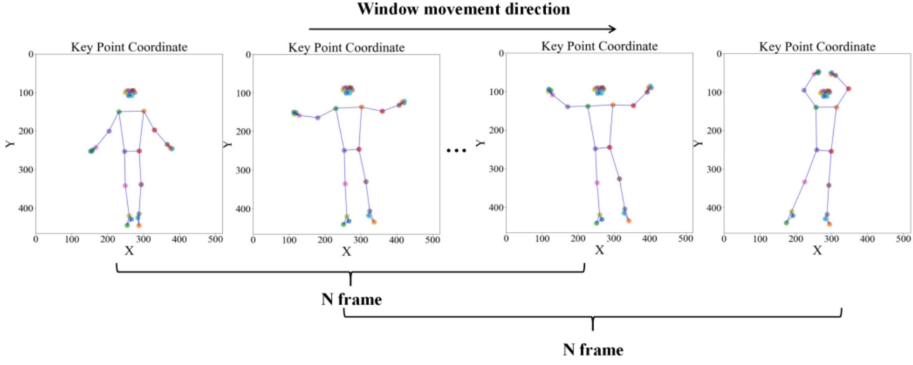


Fig. 5. Moving process of sliding window algorithm

The instantaneous velocity vector $\mathbf{v}_t^j \in \mathbb{R}^2$ of joint j at frame t is computed using the central difference method:

$$\mathbf{v}_t^j = \frac{\hat{p}_{t+1}^j - \hat{p}_{t-1}^j}{2\Delta t} \quad (8)$$

Where, $\hat{p}_t^j$ represents the normalized coordinate of joint j at frame t . The term $\Delta t = 1/f_{\text{frame}}$ denotes the inter-frame time interval, determined by the camera frame rate f_{frame} . This central difference formulation effectively mitigates phase offsets typically introduced by forward or backward difference methods. The x and y components of $\mathbf{v}_t^j$ characterize the horizontal and vertical motion rates of the joint, respectively.

Subsequently, the acceleration vector $\mathbf{a}_t^j \in \mathbb{R}^2$ is defined as the first derivative of the velocity:

$$\mathbf{a}_t^j = \frac{\mathbf{v}_{t+1}^j - \mathbf{v}_{t-1}^j}{2\Delta t} \quad (9)$$

As the first derivative of the velocity vector, the acceleration $\mathbf{a}_t^j$ quantifies the instantaneous rate of change in the motion state of joint j . This metric is inherently sensitive to abrupt directional shifts and speed fluctuations, thereby effectively capturing the jerky or unstable motor patterns often associated with neuromuscular deficits.

To enhance the robustness of action discrimination, three distinct types of motion features are aggregated within a sliding window of size N .

First, the velocity scalar $s_k^j = \|\mathbf{v}_k^j\|_2$ is calculated for each joint in the window's central frame k . This utilizes the L^2 norm to eliminate direction dependence, focusing solely on movement speed.

Second, the mean acceleration magnitude $\bar{a}_k^{(j)}$ is computed to quantify action stability. This is defined as the average L^2 norm of the acceleration vectors across the window:

$$\bar{a}_k^{(j)} = \frac{1}{N} \sum_{t=k-[N/2]}^{k+[N/2]} \|\mathbf{a}_t^j\|_2 \quad (10)$$

Where, N represents the total number of frames in the sliding window, and the summation covers the temporal span centered at k .

Third, the inter-joint coordination is measured using the cosine similarity of velocity vectors for functional joint pairs (e.g., shoulder-elbow, hip-knee):

$$\rho_k^{j_1 j_2} = \frac{\mathbf{v}_k^{j_1} \cdot \mathbf{v}_k^{j_2}}{\|\mathbf{v}_k^{j_1}\|_2 \|\mathbf{v}_k^{j_2}\|_2} \quad (11)$$

This feature $\rho_k^{j_1 j_2}$ captures the cooperative motion patterns between joints j_1 and j_2 , reflecting the synergy of the limb movement.

Finally, the temporal motion feature vector f_k^{time} is constructed by concatenating these motion descriptors for the central frame k :

$$f_k^{\text{time}} = \left[\left\{ s_k^{(j)} \right\}_{j=1}^{33}, \left\{ \bar{a}_k^{(j)} \right\}_{j=1}^{33}, \left\{ \rho_k^{(j_1, j_2)} \right\}_{(j_1, j_2) \in \mathcal{C}} \right]^\top \in \mathbb{R}^{d_t} \quad (12)$$

Here, $\mathcal{C}$ denotes a predefined set of anatomically related joint pairs (e.g., the shoulder-elbow-wrist chain), and d_t represents the total dimensionality of the temporal feature vector. This vector is then combined with the spatial geometric feature f_k^{space} (defined in Sect. 2.2.2) to form the unified spatio-temporal coupling feature vector:

where d_s is the dimension of the spatial features. This unified vector f_k serves as a comprehensive descriptor for subsequent action classification and anomaly detection⁴. The progressive update mechanism of the sliding window balances the computational demands of real-time dynamic characterization with the need for temporal context, providing a time-alignment-sensitive criterion for the improved DTW algorithm.

2.3 Improved Dynamic Time Warping Algorithm Based on Spatio-Temporal Coupling Feature Fusion

The traditional DTW algorithm has limitations in rehabilitation action evaluation: independent processing time alignment or single feature, it is difficult to capture the nonlinear dynamics of space-time interaction, resulting in alignment deviation and affecting positioning accuracy. In view of this, this paper proposes a DTW algorithm (STCF-DTW) that combines spatio-temporal coupling features. The algorithm is based on the spatio-temporal coupling feature vector $f_t = [f_t^{\text{space}}; f_t^{\text{time}}]$, embedded in the DTW framework and introduces physiological constraints to achieve innovation.

The core of the algorithm is to define the weighted spatio-temporal Euclidean distance as the inter-frame similarity measure:

$$d(f_t, f_r) = \sqrt{\alpha \cdot \|f_t^{\text{space}} - f_r^{\text{space}}\|^2 + \beta \cdot \|f_t^{\text{time}} - f_r^{\text{time}}\|^2} \quad (13)$$

Where, f_t and f_r represent the frame features of the test sequence and the reference template, respectively. The weight coefficients α and β were determined by independent validation set optimization. STCF-DTW imposes double constraints: the local slope constraint limits the maximum bending factor of the path to $r = 2$ (i.e., $|i' - j'| \leq r$), avoiding excessive distortion of the timing sequence; the joint topology constraint divides the joint points into functional groups (such as the ‘shoulder-elbow-wrist’ ‘chain’), and preferentially aligns the joints of the same functional group during calculation.

$$D(i, j) = d(f_i, f_j) + \min \left\{ D(i-1, j-1), D(i-1, j), D(i, j-1) \right\} \quad (14)$$

Where, $D(i, j)$ is the value (i, j) of the position in the cumulative cost matrix, which represents the minimum alignment cost between the i frame before the test sequence and the j frame before the reference template. $d(f_i, f_j)$ is the inter-frame distance between the feature vector f_i of the i frame of the test sequence and the feature vector f_j of the j frame of the reference sequence, which is calculated by the weighted spatio-temporal Euclidean distance (WSTED).

The STCF-DTW algorithm integrates spatio-temporal coupling features into the alignment process and ensures path anatomical rationality through physiological constraints. This fusion strategy enhances the fine - grained anomaly localization in rehabilitation action assessment. Unlike traditional DTW, which outputs overall similarity, STCF-DTW accurately identifies specific sub- actions (e.g., insufficient elbow flexion). This provides a reliable basis for generating semantic correction tips (e.g., “Increase elbow flexion”).

3 Experiments

3.1 Dataset and Experimental Setup

To validate our method, we conducted extensive experiments on the StrokeRehab dataset [17]. This dataset comprises 51 stroke patients and 20 healthy individuals. Post-stroke recovery is typically incomplete, necessitating rehabilitation that involves repetitive functional tasks and daily activities such as teeth brushing, combing, and drinking. All participants performed six types of daily actions (e.g., feeding, teeth brushing), each repeated 10 times. The dataset synchronously captures monocular RGB video and upper-limb 9-axis IMU data. Rehabilitation therapists annotated action stages and abnormal types on a frame-by-frame basis. The patient group exhibited typical motor compensation characteristics (e.g., trajectory discretization, co-movement abnormalities), while the healthy group served as a source of standard templates.

The experimental environment was configured with an AMD Ryzen 7 5800 H CPU and an NVIDIA GeForce RTX 3050 Ti Laptop GPU. The algorithm was implemented using the PyTorch 1.13.1 framework.

3.2 Evaluating Indicator

To objectively assess the improved dynamic time warping (DTW) algorithm's performance in rehabilitation action alignment and abnormal positioning, this study employs three quantitative indicators widely recognized in academia. The mathematical definitions, calculation processes, and clinical significance of these indicators are outlined below:

(1) Normalized DTW Distance (NDTW)

This index eliminates the interference of sequence length variation on similarity measurement, and objectively quantifies the consistency of the overall motion pattern of the test sequence Q and the standard template C :

$$\text{NDTW}(Q, C) = \frac{\text{DTW}(Q, C)}{\sqrt{|Q| \times |C|}} \quad (15)$$

Where, $\text{DTW}(Q, C)$ represents the minimum alignment cost calculated by the classical DTW algorithm, and $|Q|$ and $|C|$ are the number of frames of the test sequence and the standard template, respectively. The NDTW value range is $[0, +\infty)$, and the smaller the value is, the higher the action similarity is, and the NDTW of healthy subjects tends to zero under ideal conditions.

(2) Warping Degree (WD)

This index characterizes the timing distortion of the alignment path and reflects the robustness of the algorithm to the difference of action rhythm:

$$\text{WD}(P) = \frac{L_P}{\max(|Q|, |C|)} \quad (16)$$

Where, L_P is the length of the optimal path P of DTW. The theoretical minimum value of WD is 1.0 (indicating no timing distortion). The larger the value, the more the alignment path deviates from the diagonal, and the more the mapping relationship does not conform to the continuity of physiological motion.

(3) Intersection over Union (IoU)

This index evaluates the spatial and temporal accuracy of the system's positioning of abnormal action segments, and calculates the overlap ratio between the predicted abnormal period A_{pred} and the marked real abnormal period A_{gt} :

$$\text{IoU} = \frac{|A_{\text{pred}} \cap A_{\text{gt}}|}{|A_{\text{pred}} \cup A_{\text{gt}}|} \quad (17)$$

Here, the notation $|\cdot|$ denotes the duration of the time interval (measured in frames). The IoU metric ranges within $[0, 1]$, where a higher value indicates greater localization accuracy. In clinical practice, an IoU threshold of ≥ 0.7 is typically required to satisfy the precision standards for rehabilitation guidance.

3.3 Contrast Experiment

To systematically verify the effectiveness of the proposed STCF-DTW algorithm in rehabilitation action alignment and anomaly localization, we compare it against four representative baseline methods. These include traditional Dynamic Time Warping (DTW) and its extensions, a shape-metric-based DTW, and a state-of-the-art deep learning alignment framework. The core mechanisms and implementation details for each method are as follows:

- (1) Traditional DTW [18]: This method employs the classical dynamic time warping algorithm using frame-by-frame Euclidean distance as the similarity metric. As it incorporates no feature enhancement or physiological constraints, it serves as a reference baseline to establish the lower bound of performance.
- (2) Multiscale DTW [19]: To reduce sensitivity to sequence length variations, this method introduces a multi-scale segmentation strategy via windowed downsampling. It utilizes a coarse-to-fine dual-layer alignment process with weighted distances to mitigate error accumulation caused by rhythmic differences.
- (3) ShapeDTW [20]: This approach utilizes shape descriptors to align local subsequences after mapping them into a high-dimensional feature space. By accounting for both local morphology and global structural information, it improves the capability to depict complex action morphologies.
- (4) DeepAlign [21]: Representing the current data-driven deep learning paradigm, this is an end-to-end deep alignment network based on the Transformer architecture. It learns implicit inter-frame correspondences through self-attention mechanisms and outputs a soft alignment matrix.

Table 3. Comparative experiment

Method	NDTW	WD	IoU
Traditional DTW	1.129	1.848	0.455
Multiscale DTW	0.746	1.406	0.631
ShapeDTW	0.616	1.310	0.714
DeepAlign	0.836	1.567	0.566
Proposed STCF-DTW	0.473	1.158	0.814

The best results are marked in bold font.

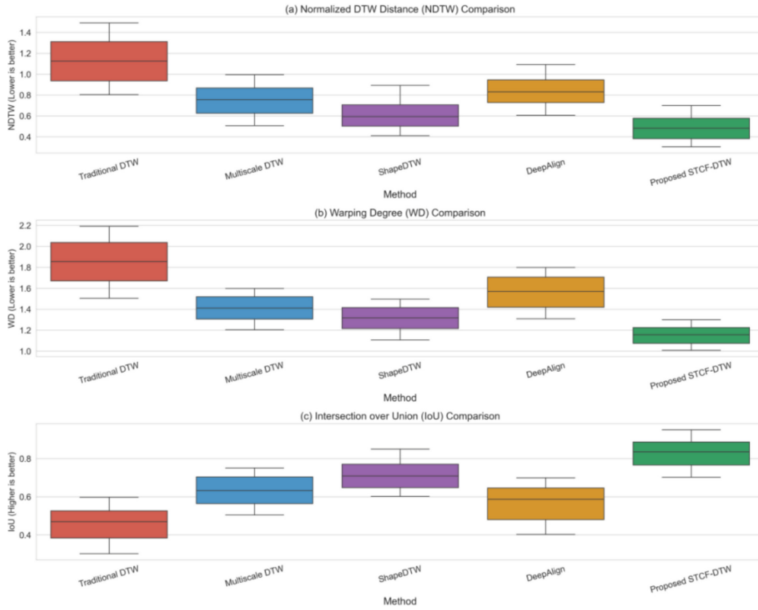


Fig. 6. Comparison of experimental visualization results (a) Normalized DTW Distance (NDTW) Comparison; (b) Warping Degree (WD) Comparison; (c) Intersection over Union (IoU) Comparison

Table 3 and Fig. 6 indicate that Traditional DTW, which relies solely on frame-by-frame Euclidean distance, fails to capture spatiotemporal coupling. Consequently, it yields high NDTW values, Warping Degree (WD) scores that deviate from the physiological range, and low IoU metrics, highlighting its susceptibility to rhythm differences and occlusions. While Multiscale DTW and ShapeDTW employ hierarchical sampling and local shape contexts, respectively, to mitigate misalignment, their reliance on single-channel features limits their capacity to describe joint coordination. As reflected in Fig. 6, this limitation leads to elevated box plots with divergent whiskers, indicating higher error rates and significant variance. DeepAlign relies on data-driven implicit alignment; however, when training and clinical distributions differ, attention drift occurs, resulting in a dispersed IoU distribution. In contrast, STCF-DTW fuses spatial geometry, velocity, and acceleration features. It further employs joint-function groups as topological constraints to compress the search space and suppress non-physiological distortions. As shown in Fig. 6, this approach significantly reduces metric values (lower boxes) and minimizes variance (converged whiskers). Table 3 confirms that STCF-DTW achieves optimal values across all corresponding columns, validating the effectiveness of spatiotemporal coupling features and physiological priors in enhancing both alignment and anomaly localization.

3.4 Ablation Experiment

To rigorously analyze the incremental contribution of each component within the proposed STCF-DTW framework, the ablation experiment establishes four algorithm variants under unified data partitioning and evaluation protocols. First, as a baseline, Traditional DTW relies solely on frame-by-frame Euclidean distance to measure sequence similarity. Second, a multi-scale weighting mechanism (+Multi-scale) is introduced to adaptively weight inter-frame distances using a coarse-to-fine dual-layer strategy, thereby mitigating errors caused by rhythmic variations. Third, joint topology constraints (+Topology), based on anatomical functional groups, are added to restrict the alignment path to a physiologically feasible region via local slope thresholds. Finally, the aforementioned constraints are deeply integrated with spatio-temporal coupling features to form the complete STCF-DTW (Full Model), which possesses comprehensive representation, alignment, and interpretation capabilities. This progressive design provides a controlled experimental framework for quantifying both the independent contributions and the synergistic effects of each module.

Table 4. Ablation experiment

Method	NDTW	WD	IoU
Baseline	1.129	2.048	0.429
+ Multi-scale	0.796	1.606	0.581
+ Topology	0.629	1.410	0.714
Full Model	0.465	1.160	0.861

The best results are marked in bold font.

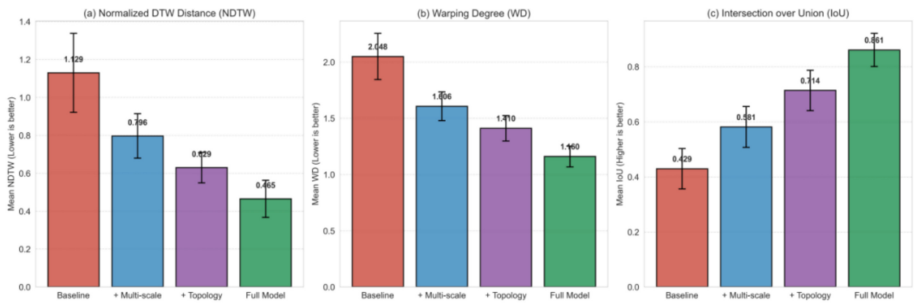


Fig. 7. Visualization results of ablation experiment (a) Normalized DTW Distance (NDTW); (b) Warping Degree (WD); (c) Intersection over Union (IoU)

Table 4 and Fig. 7 illustrate a progressive performance improvement driven by the “representation-constraint-fusion” strategy. The baseline model relies solely on frame-by-frame Euclidean distance, rendering the alignment path highly sensitive to rhythmic disturbances; consequently, it exhibits high alignment costs, significant temporal distortion, and imprecise anomaly localization. Upon introducing multi-scale weighting,

hierarchical sampling anchors the global structure at a coarse-grained level before refining local details at a fine-grained level. This results in a reduction of the overall alignment cost by approximately 30%, a simultaneous convergence of distortion amplitude, and a significant initial improvement in positioning accuracy. Subsequently, joint topology constraints are applied to embed prior path costs based on anatomical functional groups. This significantly suppresses non-physiological warping, further reducing alignment costs and distortion while boosting positioning accuracy by over 20%. Finally, the integration of spatio-temporal coupling features enriches the kinematic context, ensuring that the cost matrix reflects tighter topological and dynamic consistency, thereby allowing all three metrics to reach optimal levels. As shown in Fig. 7, the gradual decrease in bar height and the significant narrowing of error bars fully confirm the complementary synergy and statistical robustness of the proposed triple-constraint.

3.5 Semantic Correction Hints

During actual system testing, the model not only outputs quantitative metrics for action alignment and anomaly localization but also synchronously generates clinically oriented auxiliary insights. These include the “Joint Anomaly Assessment” (Table 5), “Movement Phase Assessment” (Table 6), and “Semantic Correction Suggestions” (Table 7). Collectively, these tables detail the maximum anomaly value and severity level for each joint, along with the specific time windows of detected issues. By summarizing average abnormalities, primary defects, and semantic correction suggestions according to exercise stages, the system provides patients with an actionable basis for improvement immediately following training.

Table 5. Joint Anomaly Assessment

Joint	Max Anomaly	Severity	Problem Period
R. Shoulder	0.388	Normal	
R. Elbow	0.912	Severe	30%–50%
R. Wrist	0.388	Normal	
L. Shoulder	0.870	Severe	40%–60%
L. Elbow	0.355	Normal	
L. Wrist	0.372	Normal	
R. Hip	0.389	Normal	
R. Knee	1.000	Severe	50%–70%
R. Ankle	0.375	Normal	
L. Hip	0.370	Normal	
L. Knee	0.359	Normal	
L. Ankle	0.389	Normal	

Table 6. Movement Phase Assessment

Movement Phase	Avg Anomaly	Severity	Main Issue
0–20%	0.191	Normal	
20–40%	0.208	Mild	R. Elbow flexion deficiency
40–60%	0.272	Mild	L. Shoulder elevation deficiency
60–80%	0.282	Mild	R. Knee extension deficiency
80–100%	0.189	Normal	

Table 7. Semantic Correction Suggestions

Correction Suggestion	Priority	Expected Effect
R. Elbow flexion deficiency: Increase flexion angle by 15° during mid-cycle (30%–50%)	High	Improve upper limb extension
L. Shoulder elevation deficiency: Increase elevation height by 8 cm during mid-late cycle (40%–60%)	High	Enhance shoulder mobility
R. Knee extension deficiency: Improve extension by 12° during late cycle (50%–70%)	High	Enhance lower limb coordination
Overall movement rhythm too fast: Reduce execution speed by 20%	Medium	Optimize movement rhythm

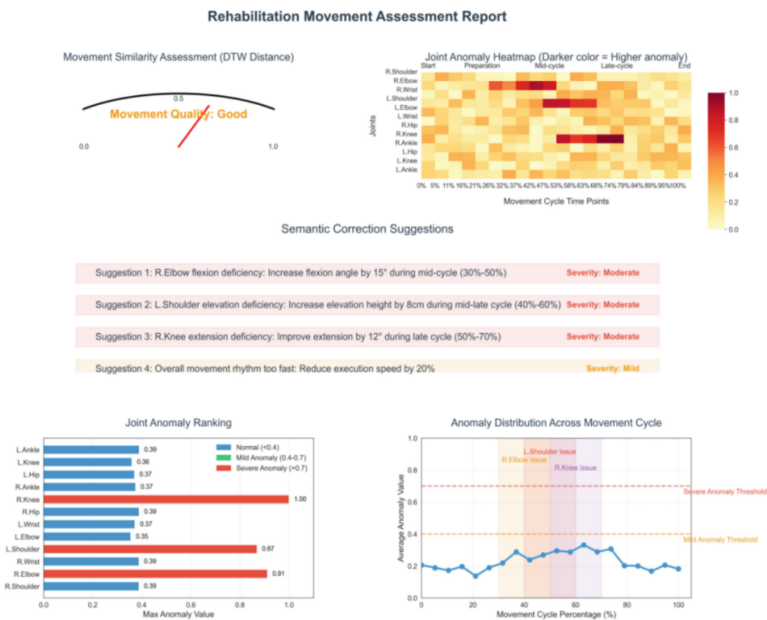


Fig. 8. Semantic correction hints (a) Movement Similarity Assessment (DTW Distance); (b) Joint Anomaly Heatmap (Darker color = Higher anomaly); (c) Semantic Correction Suggestions; (d) Joint Anomaly Ranking; (e) Anomaly Distribution Across Movement Cycle

To present evaluation results intuitively, the system generates a comprehensive visual report, as illustrated in Fig. 8. The upper-left dashboard displays overall action similarity, while the adjacent upper-right heatmap visualizes joint abnormalities, with color intensity reflecting severity. The central section organizes prioritized semantic corrective prompts into actionable cards. Below, the lower-left bar chart ranks specific joint abnormalities by severity level. Finally, the lower-right timeline visualizes anomaly distribution and threshold intervals across the entire action sequence, facilitating rapid localization of problematic temporal windows.

Through this dual-channel “text + graphics” output mechanism, patients can instantly identify key performance defects and define their spatiotemporal distribution. Guided by these combined textual and visual prompts, users can perform targeted self-adjustments. Ultimately, this enables a high-frequency, precise, and home-based rehabilitation feedback loop without requiring additional specialized equipment.

4 Conclusion

This paper proposes an intelligent rehabilitation action evaluation system that integrates improved Dynamic Time Warping (DTW) with deep learning. To address the limitations of existing methods regarding interpretability, anomaly localization accuracy, and robustness, we construct a collaborative framework driven by skeleton coordinates that synergizes spatiotemporal coupling feature extraction with multi-scale weighted DTW. Experimental results on the StrokeRehab dataset demonstrate that the proposed system significantly outperforms both traditional DTW and various deep learning baselines. Specifically, it achieves superior performance across three key metrics: Normalized DTW distance (NDTW), Warping Degree (WD), and Intersection over Union (IoU), thereby validating its practicality and stability in complex rehabilitation scenarios.

Furthermore, the system transforms anomaly localization results into a closed-loop “score-localization-interpretation” feedback mechanism via a semantic correction module. This empowers patients to engage in high-frequency, precise self-training within unsupervised environments. Ablation studies confirm that the synergistic integration of the multi-scale weighting mechanism, joint topology constraints, and spatiotemporal coupling features is crucial for fine-grained motion evaluation.

Future work will explore multimodal data fusion (e.g., incorporating IMU and depth map data) to further enhance robustness against occlusion. Additionally, we plan to incorporate a reinforcement learning mechanism to enable adaptive, dynamic training programs. These initiatives aim to advance the development of intelligent and personalized home rehabilitation solutions.

References

1. Grefkes, C., Fink, G.R.: Recovery from stroke: current concepts and future perspectives. *Lancet Neurol.* **19**(8), 660–671 (2020)
2. Panwar, M., et al.: Rehab-net: deep learning framework for arm movement classification using wearable sensors for stroke rehabilitation. *IEEE Trans. Biomed. Eng.* **66**(11), 3026–3037 (2019)

3. Sengupta, N., Rao, A., Yan, B., Palaniswami, M.: A survey of wearable sensors and machine learning algorithms for automated stroke rehabilitation. *IEEE Access*. **12**, 36026–36054 (2024)
4. Dibenedetto, D., Sotiropoulos, S., Polignano, M., Cavallo, G., Lops, P.: Comparing human pose estimation through deep learning approaches: an overview. *Comput. Vis. Image Underst.* **252**, 104297 (2025)
5. Rosa, B., Colombo Zefinetti, F., Vitali, A., Regazzoni, D.: RGB-D sensors as marker-less MOCAP systems: a comparison between microsoft kinect V2 and the new microsoft kinect azure. *Adv. Intell. Syst. Comput.* **264**, 359–367 (2021)
6. Aulton, C., Wakili, L., Strafford, B.W., Davids, K., Chiu, C.Y.: The application of deep learning human pose estimation in sport: a systematic review. *Sports Med. Open.* **11**(1), 155 (2025)
7. Stenum, J., Rossi, C., Roemmich, R.T.: Two-dimensional video-based analysis of human gait using pose estimation. *PLoS Comput. Biol.* **17**(4), e1008935 (2021)
8. Yu, X., Xiong, S.: A dynamic time warping based algorithm to evaluate kinect-enabled home-based physical rehabilitation exercises for older people. *Sensors*. **19**, 2882 (2019)
9. Deb, S., Islam, M.F., Rahman, S., Rahman, S.: Graph convolutional networks for assessment of physical rehabilitation exercises. *IEEE Trans. Neural Syst. Rehabil. Eng.* **30**, 410–419 (2022)
10. Wang, J., et al.: Automatic rehabilitation exercise task assessment of stroke patients based on wearable sensors with a lightweight multichannel 1D-CNN model. *Sci. Rep.* **14**, 19204 (2024)
11. Sardari, S., Sharifzadeh, S., Daneshkhah, A., Nakisa, B., Loke, S.W., Palade, V.: Artificial intelligence for skeleton-based physical rehabilitation action evaluation: a systematic review. *Comput. Biol. Med.* **158**, 106835 (2023)
12. Wang, S., Huang, Q., Zhang, Y., Li, X., Nie, Y., Luo, G.: Review of action recognition based on multimodal data. *J. Image Graphics.* **27**(11), 3139–3159 (2022)
13. Chen, Z., Li, S., Yang, B., Li, Q., Liu, H.: Multi-scale spatial temporal graph convolutional network for skeleton-based action recognition. *Proc. AAAI Conf. Artif. Intell.* **35**(2), 1113–1122 (2021)
14. Liu, Y., Zhang, Y.A., Zeng, M., Zhao, J.: A novel distance measure based on dynamic time warping to improve time series classification. *Inf. Sci.* **656**, 119921 (2024)
15. Tang, Y., Wang, Z., Lu, J., Feng, J., Zhou, J.: Multi-stream deep neural networks for RGB-D egocentric action recognition. *IEEE Trans. Circuits Syst. Video Technol.* **29**(10), 3001–3015 (2019)
16. Wang, P., Li, W., Ogunbona, P., Wan, J., Escalera, S.: RGB-D-based human motion recognition with deep learning: a survey. *Comput. Vis. Image Underst.* **171**, 118–139 (2017)
17. Kaku, A., et al.: StrokeRehab: a benchmark dataset for sub-second action identification. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*, vol. 122, pp. 1671–1684 (2022)
18. Rakthanmanon, T., et al.: Searching and mining trillions of time series subsequences under dynamic time warping. In: *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12)*, pp. 262–270 (2012)
19. Shokoohi-Yekta, M., Hu, B., Jin, H., Wang, J., Keogh, E.: Generalizing DTW to the multi-dimensional case requires an adaptive approach. *Data Min. Knowl. Disc.* **31**(1), 1–31 (2017)
20. Zhao, J., Itti, L.: ShapeDTW: shape dynamic time warping. *Pattern Recogn.* **74**, 171–184 (2018)
21. Kowalski, M., Naruniec, J., Trzcinski, T.: Deep alignment network: a convolutional neural network for robust face alignment. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2034–2043 (2017)

Author Index

C

Cai, Yanan 49, 70
Chen, Julong 95
Chen, Xiaofei 212
Chen, Yajun 82
Chen, Zhihui 165
Cuan, Yumeng 82

D

Duan, Ping 336
Duan, Ting-ting 336
Duan, Xiao-xiao 301, 336

F

Fu, Zhanbao 236

G

Gan, Yongjun 255
Gao, Xuefei 165
Gao, Yunxuan 126
Guo, Rongyan 110
Guo, Weilun 275

H

Han, Yizhou 255
He, Weize 70
Hou, Feng-Hsiung 313
Hua, Rui 95

J

Jiang, Xin 325
Jin, Meng 245

L

Li, Jun 236
Li, Shaojie 153
Li, Teng 245
Li, Weimin 275
Li, Yisong 176
Li, Zhen 95

Li, Zhongwei 325
Liang, Yujia 49
Liu, Tianyi 225
Lu, Guoyang 212
Lu, Haonan 255
Luo, Chen 95

M

Ma, Zijia 325
Mou, Xuepeng 95
Mu, Xinxing 275

P

Peng, Ling 225
Ping, Liang 212

Q

Qi, Ying 236

T

Tian, Peizhi 110

W

Wang, Bin 95
Wang, Jianlin 165
Wang, Jinyou 15
Wang, Ruiqi 126
Wang, Shiqi 126
Wang, Wei 95
Wang, Xiaotong 15
Wang, Yafei 110
Wang, Yixuan 38
Wang, Yudi 70
Wang, Yuhang 3, 15, 61
WangXin, Haoyun 61
Wei, Hao 188
Wen, Xing 165
Wu, Jiafeng 38

X

Xia, Jiangtao 245
 Xie, Linqi 245
 Xin, Xuefeng 349
 Xu, Mingyang 200
 Xu, Yijia 255
 Xue, Yufei 138

Y

Yan, Tie 126
 Yang, Changyong 212
 Yang, Luying 138
 Yang, Shengli 225
 Yang, Shiping 95
 Ye, Liang 275
 Yuan, Chao 27

Z

Zeng, Zishan 38
 Zhang, Dongmeng 138
 Zhang, Qingyin 49
 Zhang, Shangduo 245
 Zhang, Ting 288
 Zhang, Wei 3, 15, 61, 82
 Zhang, Wenya 61
 Zhang, Xiangyang 236
 Zhang, Yuheng 165
 Zhao, Liuqi 165
 Zhong, Kai 325
 Zhou, Qingying 82
 Zhou, Tong 301
 Zhu, Hua 27
 Zhu, Yongqing 95