

3D-Yoga: A 3D Yoga Dataset for Visual-based Hierarchical Sports Action Analysis

Jianwei Li¹, Haiqing Hu¹, Jinyang Li¹, and Xiaomei Zhao²

¹ Beijing Sports University, Beijing 100084, P.R. China

² Shandong Jianzhu University, Jinan 250101, P.R. China

{jianwei, hhq, ljy}@bsu.edu.cn, zhaoxiaomei20@sdjzu.edu.cn

Abstract. Visual-based human action analysis is an important research topic in the field of computer vision, and has great application prospect in sports performance analysis. Currently available 3D action analysis datasets have a number of limitations in sports application, including the lack of special sports actions, distinct class or score labels and variety of samples. Existing researches mainly use various special RGB videos for sports action analysis, but analysis with 2D features is less effective than 3D representation. In this paper, we introduce a new 3D yoga pose dataset (3D-Yoga) with more than 3,792 action samples and 16,668 RGB-D key frames, collected from 22 subjects performing 117 kinds of yoga poses with two RGB-D cameras. We have reconstructed 3D yoga poses with sparse multi-view data and carried out experiments with the proposed cascade two-stream adaptive graph convolutional neural network (Cascade 2S-AGCN) to recognize and assess these poses. Experimental results have shown the advantage of applying our 3D skeleton fusion and hierarchical analysis methods on 3D-Yoga, and the accuracy of Cascade 2S-AGCN outperforms the state-of-the-art methods. The introduction of 3D-Yoga will enable the community to apply, develop and adapt various methods for visual-based sports activity analysis.

Keywords: Yoga pose· Action analysis· 3D dataset· Human motion capture.

1 Introduction

Human action analysis is an important and challenging problem in the field of computer vision, and in recent years it has been widely applied in intelligent sports. Intelligent sports action analysis can help to improve the athletes' competitive ability or promote public scientific fitness, and usually use human motion capture (Mocap) technology to obtain 3D human movements. The traditional inertia and optical motion capture systems can track and record human motion well, but these systems need to bind sensors or paste marker points on human body which may affect human motion and have not been popularized in public exercise. Therefore, visual-based markerless motion capture technologies have being increasingly researched and used in human sports analysis.

In recent years, with the development of deep learning and large human action datasets, visual-based human action analysis has made remarkable progress. Human 3.6M [9] and NTU RGB+D [19] are two well known large scale 3D datasets, which are the potential resources for deep learning methods for human regular action recognition. Human 3.6M has 3.6 million human poses and corresponding RGB+D images,

acquired by recording the performance of 11 subjects under 4 different viewpoints. NTU RGB+D 120 contains 114,480 video sequences and over 8 million frames, and has 120 action classes performed by 106 subjects captured from 155 views with Kinect cameras. To meet the increasing application requirements in intelligent sports, some human professional sports datasets (HPSDs) also have been presented in recent years. However, the data of most HPSDs are RGB images or videos collected from the Internet [16, 23, 29, 34], and no 3D skeleton is provided. As the current mainstream method, deep learning needs large-scale human motion data to train a better model, which limits their effects on sports action analysis. The performance of algorithms of motion scoring or quality evaluation for sports training is still below the current application requirements. Especially for home fitness, such as yoga, the popularity rate is relatively high but lacking scientific guidance and feedback. Besides, a main challenge in sports action analysis is the accurate recovery of 3D human motion, and how to obtain accurate 3D human pose and analyze human action with limited information (such as data missing caused by occlusion) need to be further studied.

To address above issues, this paper proposes a hierarchical 3D yoga pose dataset called 3D-Yoga, which consists of 117 kinds of poses with 3,792 action samples, and each sample includes RGB-D image, human skeleton, pose label and quality score. Compared with existing sports action datasets, 3D-Yoga has several appealing properties: i) First, poses in 3D-Yoga are actually quite complex and challenging for 3D human pose estimation; ii) Second, data in 3D-Yoga is manually corrected and originally organized with hierarchical classification labels and pose quality scores; iii) Third, 3D-Yoga can be applied both to visual-based action recognition and action quality assessment tasks. To the best of our knowledge, 3D-Yoga is the first 3D sports dataset which covers actions task types including action recognition and quality assessment, and provides the corresponding RGB-D images and 3D skeleton. Based on 3D-Yoga, we propose a 3D skeleton fusion method to alleviate the occlusion problem and a hierarchical action analysis method for complex yoga poses. In summary the main **contributions** of this work are the followings:

- A new 3D sports action dataset with 117 categories of yoga poses performed by 22 subjects in various indoor environments;
- A sparse multi-view data alignment method to reconstruct 3D yoga poses to solve the severe self-occlusion problem;
- A hierarchical sports action analysis method through a cascade graph convolutional neural network for yoga pose classification and assessment.

2 Related work

2.1 Skeleton-based action recognition

Human action recognition (HAR) aims to identify what the action is, including action detection and action classification. Deep learning is currently the mainstream method for skeleton-based action recognition, where the most widely used models are RNNs and CNNs. RNN-based methods [5, 6, 19, 20, 31, 38] usually model the skeleton data as a sequence of the coordinate vectors along both the spatial and temporal dimensions,

where each vector represents a human body joint. CNN-based methods [12, 14, 18, 33] generally model the skeleton data as a pseudo image based on the manually designed transformation rules. Both of the RNN-based and CNN-based methods fail to fully represent the structure of the skeleton data because the skeleton data are naturally embedded in the form of graphs rather than a vector sequence or a 2D grid. In recent years, graph convolutional networks (GCNs), which generalize convolution from image to graph, have been successfully adopted in many applications. ST-GCN [36] proposes a dynamic skeleton model which can automatically learn both the spatial and temporal patterns from images, and demonstrated to be effective in learning both spatial and temporal dependencies on skeleton graphs. Thus many improvements based on ST-GCN have emerged, such as ST-TR [28], 2S-AGCN [30], CTR-GCN [3], and so on. ST-TR models dependencies between joints using a spatial self-attention module to understand intra-frame interactions between different body parts and a temporal self-attention module to model inter-frame correlations. 2S-AGCN improves the ST-GCN method by splitting the adjacency matrix representing action features into three, already containing richer behavioral information. That includes both the first and second features of the skeleton data, which represent the joint coordinates and length and direction of the bone, respectively. CTR-GCN dynamically learns different topologies and effectively aggregates joint features in different channels for skeleton-based action recognition. For sports action recognition, Li et al. [15] introduce an efficient fitness action analysis based on 2D spatio-temporal feature encoding, which can be applied in artificial intelligence (AI) fitness system. Aifit [7] introduces an automatic 3D human-interpretable feedback models for fitness training. HDVR [8] proposes a hierarchical dance video recognition framework by estimating 3D human pose from the corresponding 2D human pose sequences, but the accuracy is limited because of the lack of ground-truth 3D annotations for training the proposed semi-supervised method.

2.2 Visual-based action quality assessment

Human action quality assessment (AQA) aims to automatically quantify the performance of the action or to score its performance. General methods deployed to compare action sequences are based on estimating the distance error or dynamic time warping. Deep learning methods for AQA can be divided into RGB video-based methods and skeleton-based methods. Algorithms based on RGB video generally extract features directly from images through deep learning models, such as C3D [26], I3D [1], and TSN [35], and then extract time domain features by LSTM, pooling, and so on. The final score prediction is performed by a fully connected neural network. ScoringNet [17] and SwingNet [23] are all based on such methods, and support fine-grained action classification and action scoring. These methods mainly focus on the visual activity information of the whole scene including the performer's body and background, but ignore the detailed joint interactions. Skeleton-based methods firstly detect the human skeleton in the image or video, and then model the correlation information between human joints, so as to realize human motion modeling and motion quality evaluation. Pan et al. [24] propose to learn the detailed joint motion based on the joint relation, which consists of a joint commonality module modeling the general motion for certain body parts and a joint difference module modeling the motion differences within body parts. These

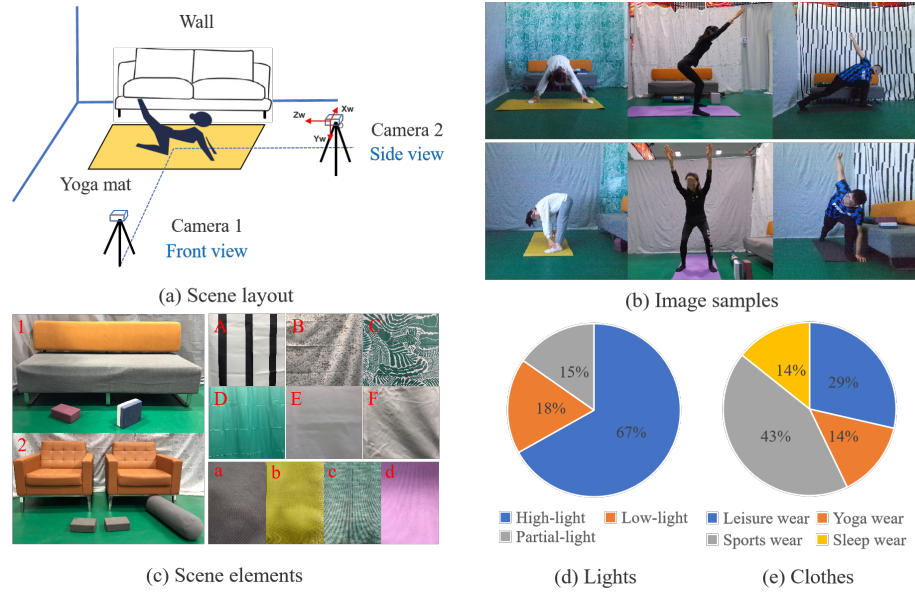


Fig. 1: The capturing of yoga poses in 3D-Yoga. (a) Scene layout. (b) Samples of yoga pose. (c) Scene elements: 1 and 2 are two indoor scenes; A, B, C, D, E, and F are the textures of the wall; a, b, c, and d are the textures of yoga mats. (d) Proportions of three light types. (e) Proportions of four cloth types.

methods can better evaluate the motion information of human body when the skeleton joint is accurate and have good interpretability. SportsCap [2] introduces a multi-stream ST-GCN method to predict a fine-grained semantic action attributes, and adopts a semantic attribute mapping block to assemble various correlated action attributes into a high-level action label for the overall detailed understanding of the whole sequence, so as to enable various applications like sports action assessment or motion scoring. However, the athletes are often in very unusual postures (such as folding and bending), which leads to poor effect of human skeleton model on 2D sports data.

2.3 Sports action dataset

UCF-sport [32] is the first sports action dataset, and contains close to 200 action video sequences collected from various sports which are typically featured on broadcast television channels such as BBC and ESPN. Since then a number of sports motion datasets [16, 23, 29, 34] used for action recognition have emerged. Diving48 [16] is a fine-grained 2D dataset for competitive diving, consisting of 18K trimmed video clips of 48 unambiguous dive sequences. Each dive sequence is defined by a combination of takeoff, movements in flight, and entry. GolfDB [23] is a 2D dataset created for general recognition applications in the sport of golf, and specifically for the task of golf swing sequencing. FineGym [29] provides coarse-to-fine annotations both temporally and semantically for gymnastics videos. There are three levels of categorical labels, and the temporal dimension is also divided into two levels, i.e., actions and sub-actions.

Yoga-82 [34] is a hierarchical 2D dataset for large-scale yoga pose recognition with 82 classes. Each image contains one or more people doing the same yoga pose, and the picture information is complex, involving different backgrounds, different clothes and colors, and different camera view angles. Pose tutor [4] curates two other fitness datasets: Pilates-32, and Kungfu-7 datasets, and combines vision and 2D pose skeleton models in a coarse-to-fine framework to obtain pose class predictions. Some 2D datasets used for AQA task also have been proposed, such as MIT Olympic sports dataset [27], Nevada Olympic sports dataset [26], and AQA-7 [25], which predict the sports performance of diving, gymnastic vault, etc. FSD-10 [21] is a figure skating dataset for fine-grained sports content analysis, which collects 1,484 clips from the worldwide figure skating championships of 10 different actions in females and males programs. Existing sports action datasets used for AQA are mainly based on publicly available RGB images or videos, but have no 3D human skeleton pose. Action analysis methods with them may focus on the global texture features and tend to ignore the motion relationship within the human skeleton joints.

3 3D Yoga dataset











In this section, we propose a new yoga pose dataset 3D-Yoga. We will introduce 3D-Yoga from data capturing, pose classification, pose assessment and data organization. The skeleton data will be made publicly available at <https://3DYogabsu.github.io>.

3.1 Yoga pose capturing

For yoga pose capturing, we deploy 22 subjects to perform yoga actions on the *daily yoga* (<https://www.dailyyoga.com.cn>) in two indoor scenes. The yoga poses are captured by two Microsoft Kinect Azure cameras from front view and side view simultaneously. As illustrated in Fig. 1 (a), two cameras are mutually orthogonal in each scene. The distances between two cameras and the center of yoga mat are 250 cm and 270 cm, respectively. Both cameras are 70 cm above the ground. Fig. 1 (b) shows some image samples in 3D Yoga dataset. The top shows images captured from the front view, while the corresponding images collected from the side view are shown at the bottom.

To achieve a variety of data samples, environments of the scenes are changed during the capturing. In total, there are five backgrounds, three ambient lights, and four yoga mats. These scene elements are shown in Fig. 1 (c): 1-2 are two indoor scenes, A-F are the textures of the wall, and a-d are the textures of yoga mates. Fig. 1 (d) shows proportions of high-light (67 %), low-light (18 %) and polarized light (15 %) (light source on the side). 22 subjects (7 male and 15 female), ranging in age from 18 to 50, are asked to perform 158 consecutive actions in *beauty back build plan*, *relaxation sleep plan* and *menstrual period conditioning plan*. Each subject has different BMI and yoga skill, and wears various styles of clothes. Fig. 1 (e) shows proportions of leisure wear (29 %), yoga wear (14 %), sports wear (43 %) and sleep wear (14 %). More specific information of the subjects is shown in the supplementary material.







Table 1: Design of the two-level hierarchical classification for yoga poses.

Classification I	Descriptions	# Labels	Examples
I. Dynamic pose	Consisted by two or more single forms of convergence.	1-4	
II. Sitting	The body pose with the pelvic bottom as the main support.	5-15	
III. Inversion	Inverted spine supported by the head, shoulder neck and upper limb.	16,17	
IV. Standing	Always based on foot as the main support type.	18-25	
V. Revolve	Spine move along its vertical axis.	26-35	
VI. Prone pose	Action supported by one of the complete surfaces of the body, or by multiple parts of one surface.	36-49	
VII. Support	Body is detached from the land, and the pose is mainly supported by the hand, elbow and foot.	50-57	
VIII. Balance	Maintain balance by regulating the limbs.	58-69	
IX. Bending	The body along a certain direction of folding or folding trend of action.	70-109	
X. Kneeling	Action supported by knee, shank and foot, belongs to kneeling type.	110-117	

3.2 Pose classification

There are 158 yoga movements performed by each subject following four exercise sets in *daily yoga*. Since some yoga poses are repeated, such as *hunker pranayama* appearing three times, we merge the same movements. The final categories of yoga poses in 3D-Yoga are adjusted to 117, and each pose is different but covered all yoga formulas. As shown in Table 1, we design a two-level hierarchy to organize these 117 categories of yoga poses, in which the first level has 10 categories (listed in the first column) and the second level are the sub-categories (labeled in the third column) of yoga poses. For the first level classification, we provide their names, detailed definitions, corresponding labels of second-level, and posture examples. The second level classification is the detailed division of first level classification, which is described in the supplementary material with labels and specific pose names.

Table 2: Scoring examples for three subjects performing two different yoga poses. S denotes strength, B denotes balance, and P&T denotes pliable and tough.

Examples	Front view	Side view	Scores (S, B, P&T)
Split boat pose			(3, 3, 3)
Split boat pose			(2, 3, 1)
Easy warrior III pose			(3, 3, 1)

3.3 Pose assessment

To provide a benchmark with domain knowledge for yoga pose quality evaluation, three yoga coaches with rich experiences (having coach cards) have completed the yoga scoring for each subject. Each sample is evaluated with two indicators, i.e., difficulty coefficient and completion score. The difficulty coefficient score is obtained by referring the standard of fitness yoga posture released on the national health yoga steering committee. The completion scoring standard is made by the coach and has four levels (0-3). The distribution of the completion score ranges from 0 to 3 in terms of strength (S), balance (B), pliable and tough (P&T). Table. 2 gives scoring examples for three subjects performing two yoga poses, i.e., *split boat pose* and *easy warrior III pose*. The detail criteria of yoga pose quality assessment are described in the supplementary material. Since the scoring for the same sample from different coaches may be various, we compute an average value as the completion score. The final evaluation score $Score_{pose}^{level}$ for each sample is multiplied by the completion score and the difficulty coefficient:

$$Score_{pose}^{level} = P_{level} \times C_{pose}, \quad (1)$$

where P_{level} denotes the difficulty coefficient, and C_{pose} denotes the completion score.

3.4 Data organization

There are 3,792 action samples and 16,668 key frames in 3D-Yoga, and the organization is shown in Fig. 2. Under the *Scene* directory, there are three folders: *Front*, *Side* and *Docs*. Under *Docs*, we provide file list, pose score and camera calibration information. There are 22 folders respectively for 22 subjects under the *Front* and *Side* folders. The folders of 7 male subjects are represented as M01, M02,..M07, and the folders of 15 female subjects are represented as F01, F02,..F15. The action label (A01, A02,...A10) represents the folders name of classification I, and the sub-action label (a01, a03,...a117) represents the folders name of classification II. In each sub-action folder corresponding to a motion segment, there are three sub-folders, i.e., *Color*, *Depth* and *Skeleton*.

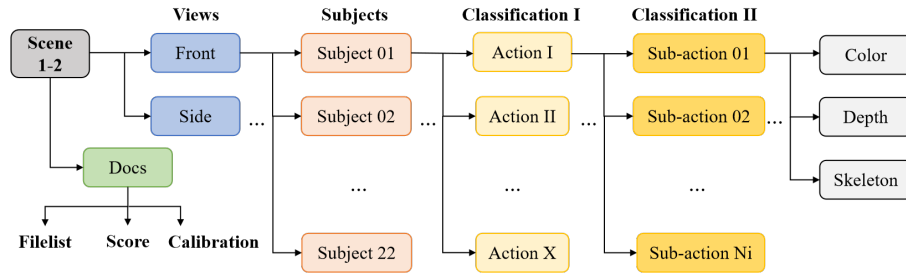


Fig. 2: Schematic representation of directory organization of the 3D-Yoga dataset.

The original unregistered full resolution depth (640×574) and color (1920×1080) images are captured by different sensors of the Kinect camera, while the skeletal joints are obtained through the bone tracking technology in Kinect SDK. In order to make the data easy applicable in many existing implementations, we provide both the color and the depth in a resolution of 1280×720 and register them both in the color camera coordinate system. Human body skeleton is composed of the three-dimensional position coordinates of the 32 main human body joints, and also is aligned in the color camera coordinate system. Skeleton data is stored in CSV format, and contains time stamps, personIDs, jointsIDs, 3D spatial coordinates and quaternions of the joints, confidences and 2D pixel coordinates of the joints.

We compare 3D-Yoga dataset with the state-of-the-art HPSDs, and Table 3 shows the comparison in terms of action classes, sample numbers, data types, sources and analysis tasks. It's obvious that 3D-Yoga is built through Mocap and provides more data type, which can be used for action recognition and AQA tasks.

4 Yoga pose analysis

The pipeline of the proposed yoga pose analysis method is shown in Fig. 3, which consists of data pre-processing and hierarchical pose analysis. The innovative points are elaborated in the following subsections.

Table 3: Comparison with the state-of-the-art professional sports datasets. R denotes action recognition task, and AQA denotes action quality assessment task.

Datasets	Classes	Samples	Data types	Sources	Tasks
UCF-sport [32]	10	150	RGB	BBC/ESPN	R
Sport-1M [11]	487	1,133,158	RGB	YouTube	R
Diving48 [16]	48	18,404	RGB	Internet	R
Yoga-82 [34]	82	28,478	RGB	Bing	R
FineGym [29]	99/288	~30K	RGB	Internet	R
FSD-10 [21]	10	1,484	RGB	YouTube	R&AQA
3D-Yoga	10/117	3,792	RGB-D+Skeleton	Mocap (Kinects)	R&AQA

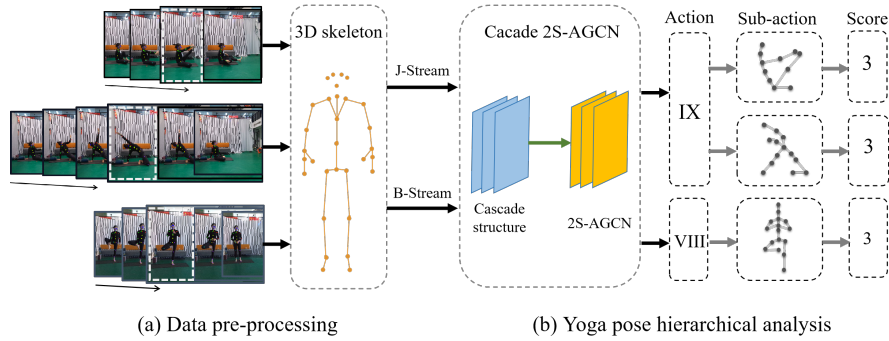


Fig. 3: The pipeline of our proposed hierarchical yoga pose analysis method.

4.1 Data pre-processing

To get a complete and accurate 3D skeleton pose for yoga pose analysis, we process the action sequences mainly through the following three steps:

Camera calibration: The calibration for Kinect cameras is performed before yoga pose capturing. More than 100 checkerboard images are selected in each calibration. The intrinsic matrix of each camera is obtained from Kinect azure SDK, while the geometric relationship of two cameras in front and side views is computed by the tool of stereo camera calibration with Matlab. The grids in the checkerboard are 10×15 , and the actual side length of each grid is 5 cm. The average re-projection error of stereo camera calibration is 2.81 pixels.

Key frame extraction: Considering that most yoga movements are static processes, we use the key frame based method to analyze yoga poses. For each captured data sequence, we first carry out coarse extraction based on image similarity, and then extract frames of each pose based on the confidence value and depth value of the human skeleton. For each yoga action, we manually select the best time synchronization key frames captured by two Kinects, and retain key frame pairs with the most obvious features. For a few dynamic yoga poses with different sequence lengths, we select multiple key frames to describe the entire movement.

3D skeleton fusion: For each pair of key frame data, the alignment of 3D skeleton pose is executed as follows:

- First, we align the 3D points from the front and side views through ICP algorithm [37] with camera calibration result as an initial value. 3D points are calculated by corresponding depth images and camera intrinsic parameter. The transformation matrix \mathbf{T}_{fs} between two views (from front to side) is obtained after ICP registration.

- Second, the 3D skeletons are fused by the transformation information, and the i -th joint \mathbf{S}_{fused}^i in fused skeleton is calculated according to the following formula:

$$\mathbf{S}_{fused}^i = \frac{\mathbf{T}_{fs}W_{fi}\mathbf{S}_{fi} + W_{si}\mathbf{S}_{si}}{W_{fi} + W_{si}}, i \in (0, 31) \quad (2)$$

where \mathbf{S}_{fi} and \mathbf{S}_{si} are the i -th joint coordinates in the front and side camera coordinate system. W_{fi} and W_{si} are the corresponding weights, determined by the joints angle and data quality, and defined as:

$$W_i = \frac{\log_4\left(\frac{c_i+1}{z_i}\right) + 1}{1 + e^{\frac{\theta_i-90}{10}}}, \theta_i \in (0, 180) \quad (3)$$

where z_i , θ_i and c_i are the depth value, joint angle and confidence coefficient respectively. The joint with high confidence and close to the camera has larger fusion weight. Eq. 3 is a refinement of Jiang et al. [10].

- Finally, the 3D skeletons are further optimized by embedding a parametric human model [22] through minimizing the following energy function:

$$\mathbf{E}_{fused}(\theta, \beta) = w_{pro}\mathbf{E}_{pro} + w_{shape}\mathbf{E}_{shape} \quad (4)$$

where \mathbf{E}_{pro} is the data term aligning 2D projections on each view to the 3D joints, \mathbf{E}_{shape} penalizes human shape prior. w_{pro} and w_{shape} are balancing weights, and set to 1 in our experiments. θ and β are two optimized parameters, which control the bone length and the joint posture respectively.

4.2 Hierarchical analysis

For yoga pose hierarchical analysis, we design the Cascade 2S-AGCN with a cascade structure and 2S-AGCN models to realize the coarse-grained to fine-grained yoga pose classification and specific yoga pose quality evaluation.

Cascade structure: Cascade 2S-AGCN contains three stages and consists of a main network and three branches (two for pose classification and one for pose assessment). The main network is modelled after AlexNet [13] and each branch is a 2S-AGCN model. The details of Cascade 2S-AGCN are described in the supplementary material. After completing the former stage by 2S-AGCN, we choose a recall rate threshold to remove error samples that will not be used to train later stages. The advantage of this filtering mechanism is that since the main and branch networks are in a unified framework, the feature maps extracted at the beginning can be shared throughout the network, rather than features being collected at each layer of the network from the original data.

Graph construction: Based on human skeleton model, we construct the skeleton information undirected graph with human skeleton link rule, in which the origin represents the key point, and the line segments represent the connection relationship of each joint

point. The input of Cascade 2S-AGCN has two streams: B-stream and J-stream. For J-stream, we use a spatio-temporal graph to simulate the structured information between them along the spatio-temporal dimensions of these joints. The structure of the graphs contains not only joint point coordinate but also spatial constraints between adjacent key points. For B-stream, the input data are the length and direction of the skeleton. We set the middle point of the pelvis as the central point, the joint near the central point as the source joint, and the joint away from the central point as the target joint. Thus the joint is the key point, the bone is the vector from one point to another, the length of the vector is the length of the bone, and the direction of the vector is the direction of the bone. The main formula for the adaptive graph convolution is as follows:

$$\mathbf{f}_{out} = \sum_k^{\mathbf{K}_v} \mathbf{W}_k \mathbf{f}_{in} (\mathbf{A}_k + \mathbf{B}_k + \mathbf{C}_k), \quad (5)$$

where \mathbf{f} denotes the feature map of the graph convolution, \mathbf{W}_k is the weight vector of the 1×1 convolution operation, \mathbf{K}_v denotes the number of subsets. \mathbf{A}_k is an $N \times N$ adjacency matrix (N denotes the number of vertexes) and represents the physical structure of the human body. \mathbf{B}_k also is an $N \times N$ adjacency matrix similar to \mathbf{A}_k , but the elements of \mathbf{B}_k are parameterized and optimized together with the other parameters in the training process. \mathbf{C}_k is a graph unique to each key frame that uses a classical Gaussian embedding function to capture the similarity between joints. The input data is in the form of $C_{in} \times T \times N$, encoded into $C_e \times T \times N$ with two embedding functions, i.e., θ and ϕ . Since the number of parameters for the convolution of 1×1 is quite large, the encoded $C_{in} \times T \times N$ is transformed into $N \times C_e T$ and $C_e T \times N$ through two conversion functions. These two matrices are multiplied to obtain the $N \times N$ similarity matrix of \mathbf{C}_k . Element $C_k^{i,j}$ in \mathbf{C}_k represents the similarity between vertex v_i and v_j . Therefore, \mathbf{C}_k is calculated by multiplying θ and ϕ :

$$\mathbf{C}_k = softmax(f_{in}^T W_{\theta k}^T W_{\phi k} f_{in}), \quad (6)$$

where T denotes the temporal length, W_{θ} and W_{ϕ} are the parameters of θ and ϕ .

Pose recognition and assessment: The yoga pose recognition task is constructed with the first and second stages of the cascade network to obtain a coarse-to-fine pose classification. For example, the *sitting leg up* and *seated butt lift back-bending* both are recognized as *Bending* at the first stage, and then identified as their respective sub-actions (label 80 and label 96) at the second stage. The yoga pose quality for each sample is evaluated with the pose completion score and difficulty coefficient. We firstly use the third stage of the cascade network to predict a completion score for each pose performed by a subject, and then multiply the difficulty coefficient score based on the yoga level. For example, if the predicted completion score of *boat pose* ($P_{level} = 3$) is 2, the final evaluation score is 6.

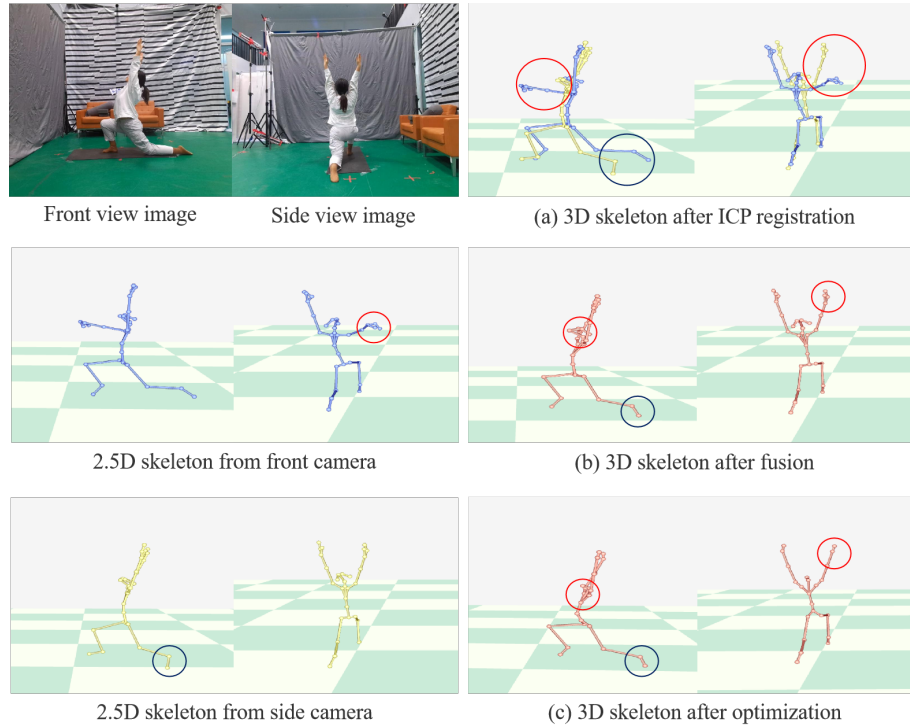


Fig. 4: Results of 3D skeleton fusion for a common yoga pose. The left column shows input data, and the right column shows output data.

5 Experiments and Analysis

In this section, we make experiments and analyses with yoga pose analysis methods on 3D-Yoga. For all experiments, we run our methods on a standard desktop PC with Intel Core i7-7700 3.6 GHz CPU.

5.1 Implementation details

Training details: As introduced in Section 4.2, there are three types of graphs in the adaptive graph convolutional block (**A**, **B**, **C**). Three-axis data is trained in the 2S-AGCN model, in which TCN and TGN are two predominant sub-models to extract time-series features and graph features from the raw data. The predicted result is the maximum classification probability of the *softmax* classifier through the *argmax* function. In our experiments, cross-entropy and AdamW are selected as the loss function and optimizer respectively, and Relu activation function is used to avoid gradient disappearance. During the training, we set the learning rate = $1e-3$, seed = 42, weight decay = 0.1, betas $\in (0.9-0.999)$, batch size = 32, and epoch = 80.

Table 4: Analysis accuracies (%) of yoga poses in 3D-Yoga. The best results are in bold.

Stages	Categories	Number	Front	Side	Combined	Fused
I	Coarse-gained	10	75.60	71.87	77.18	80.42
	Dynamic	4	77.6	68.86	78.14	73.73
	Sitting	11	74.65	69.91	85.64	87.29
	Inversion	2	100	93.48	100	92.31
	Standing	8	83.77	72.08	81.81	90.40
	Revolve	10	76.98	79.86	77.70	73.56
II	Prone	14	61.68	65.87	76.05	60.68
	Support	8	54.17	62.50	60.41	53.84
	Balance	12	82.96	82.22	80.00	92.30
	Bending	40	64.09	62.19	67.68	75.52
	Kneeling	8	60.43	69.06	66.91	69.33
	Fine-gained	117	76.10	74.13	74.92	82.94
III	Score	4	71.40	70.97	65.43	72.57

Experimental data: To compare the performance of our methods, we have carried out experiments with four types data of 3D-Yoga: 1) 2.5D skeletons of front view; 2) 2.5D skeletons of side view; 3) combined 2.5D skeletons (both of front and side views); 4) fused 3D skeleton. We use 32 joints model provided by Kinect for experiments (1-3) and 23 joints provided by the SMPL model for the experiment (4). To ensure a better training effect, we split the dataset into a training set with 16 subjects and a validation set with 6 subjects, and use hierarchical sampling in the pose labels.

5.2 Ablation studies

Yoga pose fusion: Due to human body self-occlusion, 2.5D skeleton data captured by the front or side camera always has some errors. In order to get an accurate 3D skeleton, we align 2.5D skeletons with the proposed method in Section 4.1. Fig. 4 visualizes the fused results of 3D skeleton from the front view and side view for a common yoga pose. The left column shows color images and 2.5D skeletons (in camera coordinate system) captured by the front and side Kinects, and the right column shows 3D skeletons after ICP registration, fusion and optimization respectively. The main distinctions are circled in red and blue on the Figure. The contrasts show that we have effectively restored some 3D skeleton joints after the operations of ICP registration, fusion and optimization. The data quality analysis for 3D-Yoga dataset before and after 3D skeleton fusion is presented in supplementary material.

Yoga pose analysis: Table 4 shows the quantitative analysis accuracies of yoga poses with Cascade 2S-AGCN in terms of front view skeletons, side view skeletons, combined skeletons and fused skeletons. The first line is the recognition results of Classification I in the first stage, lines 2 through 13 are the recognition results of each sub-action categories and the average accuracies of Classification II in the second stage, and the bottom line is the prediction accuracy of completion score for each yoga pose in the third stage. Number is the kinds of yoga poses in each category. Except for individual

Table 5: Comparison of recognition accuracies (%) with the state-of-the-art methods on 3D-Yoga. The best results are in bold.

Methods	Front	Side	Combined	Fused
ST-GCN [36]	53.65	56.25	56.56	58.33
2S-AGCN [30]	58.90	55.82	56.86	66.84
CTR-GCN [3]	73.02	66.24	76.40	72.93
Ours	76.10	74.13	74.92	82.94

rare poses, most recognition accuracies by using combined and fused skeletons are generally higher than the accuracies of the front view or side view skeletons. With our fused 3D skeletons, the coarse-grained recognition (Classification I) accuracy is up to 80.42 %, the fine-grained recognition (Classification II) accuracy is 82.94 %, and the prediction accuracy for pose completion score is 72.57 %. It verifies that the average performance of our yoga pose analysis method with 3D fused skeletons has been improved. More experiments results and analyses for yoga pose are provided in supplementary material.

5.3 Comparison with other methods

We also compare Cascade 2S-AGCN with the state-of-the-art methods on 3D-Yoga, and Table 5 shows comparison results (117 categories) of the recognition accuracies (%) with related methods: ST-GCN [36], 2S-AGCN [30], and CTR-GCN [3]. We run these methods on a NVIDIA RTX 1080Ti GPU, and use the same epoch equal to 80 in the comparison experiments. It can be seen that Cascade 2S-AGCN outperforms other methods and the accuracies of all methods are improved by using the fused 3D skeletons. A detailed version is provided in supplementary material.

6 Conclusions

In this work, we present a new 3D-Yoga dataset with RGB-D images, 3D skeletons, multi-level classification labels and pose scores. To evaluate 3D-Yoga, we perform a cascade graph-based convolution network to recognize coarse-grained to fine-grained yoga poses and assess the quality of each pose. Experiments have been carried out for hierarchical yoga pose analysis, and the results show that the proposed pose recognition and assessment methods have achieved a good performance. The introduction of 3D-Yoga will enable the community to apply, develop and adapt various deep learning techniques for visual-based sports activity analysis. We will further research more efficient and robust methods for sports action analysis with 3D-Yoga.

Acknowledgment This work is assisted by Siqi Wang, Rui Shi, Ruihong Cheng, Yongxin Yan and Jie Liu, five students from the school of sport engineering of Beijing Sports University, who participated in data acquisition. This work is supported by the Open Projects Program of National Laboratory of Pattern Recognition under Grant No.202100009, and the Fundamental Research Funds for Central Universities No.2021TD006.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
2. Chen, X., Pang, A., Yang, W., Ma, Y., Xu, L., Yu, J.: Sportscap: Monocular 3d human motion capture and fine-grained understanding in challenging sports videos. *International Journal of Computer Vision* **129**(10), 2846–2864 (2021)
3. Chen, Y., Zhang, Z., Yuan, C., Li, B., Deng, Y., Hu, W.: Channel-wise topology refinement graph convolution for skeleton-based action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13359–13368 (2021)
4. Dittakavi, B., Bavikadi, D., Desai, S.V., Chakraborty, S., Reddy, N., Balasubramanian, V.N., Callepalli, B., Sharma, A.: Pose tutor: An explainable system for pose correction in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3540–3549 (2022)
5. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., Darrell, T.: Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2625–2634 (2015)
6. Du, Y., Wang, W., Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1110–1118 (2015)
7. Fieraru, M., Zanfir, M., Pirlea, S.C., Olaru, V., Sminchisescu, C.: Aifit: Automatic 3d human-interpretable feedback models for fitness training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9919–9928 (2021)
8. Hu, X., Ahuja, N.: Unsupervised 3d pose estimation for hierarchical dance video recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11015–11024 (2021)
9. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **36**(7), 1325–1339 (2014)
10. Jiang, Y., Song, K., Wang, J.: Action recognition based on fusion skeleton of two kinect sensors. In: 2020 International Conference on Culture-oriented Science & Technology (ICCST). pp. 240–244. IEEE (2020)
11. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (2014)
12. Ke, Q., Bennamoun, M., An, S., Sohel, F., Boussaid, F.: A new representation of skeleton sequences for 3d action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3288–3297 (2017)
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25** (2012)
14. Li, B., Dai, Y., Cheng, X., Chen, H., Lin, Y., He, M.: Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep cnn. In: 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). pp. 601–604. IEEE (2017)
15. Li, J., Cui, H., Guo, T., Hu, Q., Shen, Y.: Efficient fitness action analysis based on spatio-temporal feature encoding. In: 2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). pp. 1–6. IEEE (2020)

16. Li, Y., Li, Y., Vasconcelos, N.: Resound: Towards action recognition without representation bias. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 513–528 (2018)
17. Li, Y., Chai, X., Chen, X.: Scoringnet: Learning key fragment for action quality assessment with ranking loss in skilled sports. In: Asian Conference on Computer Vision. pp. 149–164. Springer (2018)
18. Liu, H., Tu, J., Liu, M.: Two-stream 3d convolutional neural network for skeleton-based action recognition. arXiv preprint arXiv:1705.08106 (2017)
19. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L.Y., Kot, A.C.: Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE transactions on pattern analysis and machine intelligence* **42**(10), 2684–2701 (2019)
20. Liu, J., Shahroudy, A., Xu, D., Wang, G.: Spatio-temporal lstm with trust gates for 3d human action recognition. In: European conference on computer vision. pp. 816–833. Springer (2016)
21. Liu, S., Liu, X., Huang, G., Qiao, H., Hu, L., Jiang, D., Zhang, A., Liu, Y., Guo, G.: Fsd-10: A fine-grained classification dataset for figure skating. *Neurocomputing* **413**, 360–367 (2020)
22. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)* **34**(6), 1–16 (2015)
23. McNally, W., Vats, K., Pinto, T., Dulhanty, C., McPhee, J., Wong, A.: Golfdb: A video database for golf swing sequencing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 0–0 (2019)
24. Pan, J.H., Gao, J., Zheng, W.S.: Action assessment by joint relation graphs. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6331–6340 (2019)
25. Parmar, P., Morris, B.: Action quality assessment across multiple actions. In: IEEE winter conference on applications of computer vision (WACV). pp. 1468–1476. IEEE (2019)
26. Parmar, P., Morris, B.T.: What and how well you performed? a multitask learning approach to action quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 304–313 (2019)
27. Pirsivash, H., Vondrick, C., Torralba, A.: Assessing the quality of actions. In: European Conference on Computer Vision. pp. 556–571. Springer (2014)
28. Plizzari, C., Cannici, M., Matteucci, M.: Skeleton-based action recognition via spatial and temporal transformer networks. *Computer Vision and Image Understanding* **208**, 103219 (2021)
29. Shao, D., Zhao, Y., Dai, B., Lin, D.: Finegym: A hierarchical video dataset for fine-grained action understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2616–2625 (2020)
30. Shi, L., Zhang, Y., Cheng, J., Lu, H.: Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12026–12035 (2019)
31. Song, S., Lan, C., Xing, J., Zeng, W., Liu, J.: An end-to-end spatio-temporal attention model for human action recognition from skeleton data. In: Proceedings of the AAAI conference on artificial intelligence. vol. 31 (2017)
32. Soomro, K., Zamir, A.R.: Action recognition in realistic sports videos. *Computer vision in sports* pp. 181–208 (2014)
33. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015)
34. Verma, M., Kumawat, S., Nakashima, Y., Raman, S.: Yoga-82: a new dataset for fine-grained classification of human poses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 1038–1039 (2020)

35. Xiang, X., Tian, Y., Reiter, A., Hager, G.D., Tran, T.D.: S3d: Stacking segmental p3d for action quality assessment. In: 2018 25th IEEE International conference on image processing (ICIP). pp. 928–932. IEEE (2018)
36. Yan, S., Xiong, Y., Lin, D.: Spatial temporal graph convolutional networks for skeleton-based action recognition. In: Thirty-second AAAI conference on artificial intelligence (2018)
37. Yang, C., Medioni, G.: Object modeling by registration of multiple range images. *Image and Vision Computing* **10**(3), 145–155 (2002)
38. Zhang, P., Lan, C., Xing, J., Zeng, W., Xue, J., Zheng, N.: View adaptive neural networks for high performance skeleton-based human action recognition. *IEEE transactions on pattern analysis and machine intelligence* **41**(8), 1963–1978 (2019)