Hindawi
Computational Intelligence and Neuroscience
Volume 2022, Article ID 7500525, 9 pages
https://doi.org/10.1155/2022/7500525

Research Article

Spatial-Temporal Graph Convolutional Framework for Yoga Action Recognition and Grading

Shu Wang

School of Physical Education, Inner Mongolia Minzu University, Tongliao, Inner Mongolia 028000, China

Correspondence should be addressed to Shu Wang; wangshu2012@imun.edu.cn

Received 20 January 2022; Revised 8 February 2022; Accepted 24 February 2022; Published 29 March 2022

Academic Editor: Baiyuan Ding

The rapid development of the Internet has changed our lives. Many people gradually like online video yoga teaching. However, yoga beginners cannot master the standard yoga poses just by learning through videos, and high yoga poses can bring great damage or even disability to the body if they are not standard. To address this problem, we propose a yoga action recognition and grading system based on spatial-temporal graph convolutional neural network. Firstly, we capture yoga movement data using a depth camera. Then we label the yoga exercise videos frame by frame using long short-term memory network; then we extract the skeletal joint point features sequentially using graph convolution; then we arrange each video frame from spatial-temporal dimension and correlate the joint points in each frame and neighboring frames with spatial-temporal information to obtain the connection between joints. Finally, the identified yoga movements are predicted and graded. Experiment proves that our method can accurately identify and classify yoga poses; it also can identify whether yoga poses are standard or not and give feedback to yogis in time to prevent body damage caused by nonstandard poses.

1. Introduction

Yoga has become a very trendy fitness exercise in today's life. But yoga is much more than just a fitness exercise. Yoga is a physical and mental discipline that combines art, science, and philosophy. Yoga can help people regulate their breathing, keep their bodies healthy, and also calm their moods. In today's highly developed Internet, according to incomplete statistics, yoga has become the preferred fitness exercise for 300 million people [1]. As a scientific exercise, yoga encompasses breath control exercises, body stretching exercises, and mind cleansing [2]. Yoga originally originated in ancient India, then spread to the West, where it became a mainstream Western fitness modality, and then eventually spread globally with the Internet, becoming one of the most popular exercise cultures worldwide [3]. According to a joint UK and US survey, the demographic profile of the yoga training population found in the demographics indicates that women are the main enthusiasts of the sport, accounting for 85% of the total number of yoga practitioners [4–6].

Numerous studies have proven that yoga exercises are beneficial to the human body. There is also a large amount of research in rehabilitation on how to make yoga training work better for patients in their recovery process. This is one of the reasons why yoga has become a favorite exercise for many people [7]. In addition, research has proven that yoga has a complementary healing effect in the direction of eating disorders; it can modify the patient's eating habits and keep diet [8]. In the interviews of yoga practitioners, it was learned that yoga gave them a positive and subjective life experience, making them healthier and living an optimistic life. There were significant improvements in self-care, self-activity, life comfort, and dwelling senses [9–11]. In fact, most of the experience that yoga brings to people comes from the yoga instructor. The instructor, as the guide of yoga, influences the yoga student in an invisible way with his or her philosophy of teaching, teaching environment, outlook on life, values, and demonstration of yoga effectiveness [12].

Although some researchers have demonstrated that yoga can be practiced without differentiating between

"traditional" and "authentic" issues [13], most people currently prefer modern yoga. Modern yoga is simpler and less demanding in terms of postural alignment and breathing exercises [14]. This is one of the reasons why modern yoga has turned into a healthy exercise for young and old alike. However, due to the overall economic development, yoga has gradually become commercialized. With the commercialization, the expression of yoga has become diversified and more and more people have become attracted to yoga. In our literature research, we found that yoga is becoming a synonym for young, beautiful, and hot women [15]. Yoga can be found in various fashion magazines and shows yoga poses that have a certain ornamental quality and at the same time these poses are difficult in the eyes of professionals. For ordinary people, they are more attracted by the ornamental poses of yoga, but these poses are risky for them. Commerce has made yoga idealized in order to facilitate promotion and thus attract consumers [16]. However, the commercialization of yoga is also a double-edged sword. Consumers are likely to cause irreversible damage to their bodies in the process of blindly imitating yoga poses due to the unknown nature of the poses, which is a potential risk in yoga training.

Traditionally, yoga is taught face-to-face, with the yoga instructor instructing in person whether the yoga poses are standard or not. This kind of teaching can make yoga students have a more direct feeling of standard yoga movements. However, with the advent of the 5G era and the rapid development of short videos, short video platform bloggers often adopt online teaching methods to teach yoga poses in order to attract fans. This is also the way most people learn yoga at present. Most people choose to watch videos while imitating to achieve the purpose of learning yoga. However, most people do not have professional yoga equipment and props, and they are not clear enough about the standard yoga postures. Blindly imitating the yoga postures in the videos has a great risk of physical injury. To solve this problem, in our work, we propose to use real-time posture detection technology to detect posture movements of yoga students and then use deep learning algorithms to grade and match yoga movements. A reference movement is given to the yoga students, and for the nonstandard movements, the yoga students are prompted in time to prevent the occurrence of physical injuries. In the specific experiment, we use the deep camera to capture the training postures of yoga students and decompose the postures to understand the yoga movements from the computer level. The postures are then compared with a standard database to verify whether the postures are standardized and to give feedback to the yoga students. Experiments show that the method proposed in our research can provide effective feedback to yoga trainees on the grading of yoga poses. The contributions of this paper can be summarized as follows.

The rest of this paper is organized in the following manner. Section 2 discusses the work related to deep camera and action recognition. Section 3 introduces the skeleton recognition principle of graph convolution, then introduces the residual unit and multistream input structure, and finally introduces the optimization principle of the partial perception framework. Section 4 reports the experimental data collection, model training details, and analysis of experimental results. Finally, Section 5 concludes our research and reveals some further research work.

2. Related Work

The presentation of human motion postures in 3D space often requires the use of depth cameras. Information such as joint angles and skeletal space points can be deduced from the depth camera or the spatial position data of the human body [17]. Different poses can generate different skeletal contours, and to solve this problem, some researchers have proposed the idea of spatial segmentation, which takes an approximate mapping approach to define the location of spatial points for each segmented region. Literature [18] proposed a joint distribution method, which takes a bidirectional derivative approach to the mapping function. Literature [19] also uses the joint distribution rule, and unlike the former, the method adopts a Bayesian algorithm to obtain the image contour conditions. The final distribution of the image contour conditions will be mapped to the hybrid framework to obtain the spatial distribution features. Literature [17] additionally uses learning conditional distributions when learning features in the hybrid framework to obtain the image contour features more directly. In [20], to solve the image contour error problem caused by pose ambiguity, the researcher distributed three depth cameras into different angles to capture the human motion contour in all directions and obtained the skeletal spatial position from a standard dataset. In [21], the researcher used the SVM method to learn different pose features and perform pose prediction in the acquired 3D shape data. This proposed method links contours and 3D shapes but requires the support of large databases. For motion capture depth cameras, calibration of the depth camera is also required to ensure accuracy in 3D reconstruction work. In [22], the researcher applied the EM algorithm to calibrate the human action pose for multicamera linkage, and the mapping of 2D contours to 3D skeletal joints was achieved by training a neural network. Literature [23] adopts hybrid probabilistic PCA to predict the 3D body structure captured by the depth camera, which improves the 3D joint point coordinate accuracy.

Human motion recognition techniques originated from skeletal annotations [24, 25] by video clips [26, 27] to obtain the motion pose of each frame, which was then obtained by manual criteria. Previous human action recognition methods are based on RGB images, but this method is limited to the influence of nonobjective environments. The human skeleton-based action recognition method is less influenced by the nonobjective environment. This method can acquire the spatial-temporal features between joint points and learn the connection between features in a neural network to predict the human pose. Current neural network architectures that can be combined with the human skeleton approach are recurrent neural networks (RNN) [28, 29], long short-term memory networks (LSTM) [30, 31], convolutional neural networks (CNN) [32], etc. To make the human skeleton approach more general, [25, 26] proposed

to use the heat map as a complement to the skeleton information and to use the human pose image in each video frame for the encoding process. The feature communication between bone joint points is shown in Figure 1.

Literature [33] proposed a method to construct a human action dataset combining skeleton information with video in order to improve the pose estimation and action recognition accuracy of CNN networks. Literature [34] proposed a multitask parallel learning framework to improve the accuracy and stability of body joint detection. Literature [35] proposed a human intention algorithm aiming at learning behavioral action features through environmental assistance. Literature [36] took the approach of attention mechanism, which divides the human body into different parts and obtains attention from each part separately to recognize actions. Some researchers have found that the spatial-temporal graph convolution network (ST-GCN) can utilize the spatial-temporal information of skeletal articulation points effectively. It performs spatial-temporal convolution on the skeletal graph, models the graph representation of each skeleton, and uses a subsequent temporal filer to capture dynamic temporal information, as shown in Figure 2.

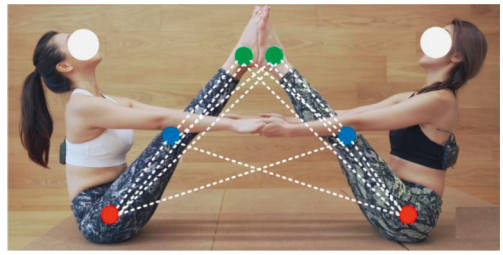


Figure 1: The feature communication between bone joint points.

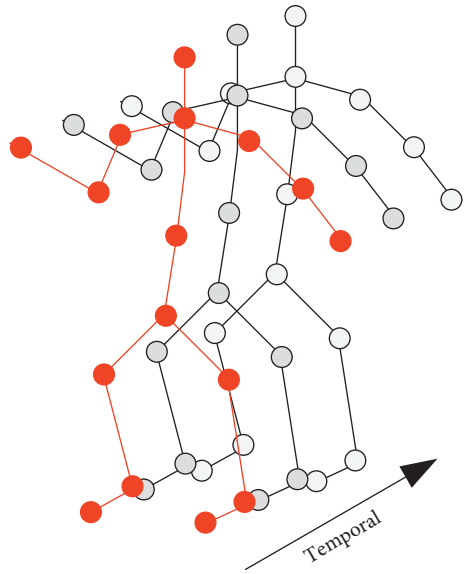


Figure 2: The principle of skeleton spatiotemporal feature extraction.

3. Method

3.1. Graph Convolutional Network. Benefiting from [37], the sequence of each frame t of the human skeleton in space is expressed as follows:

$$f_{\text{out}} = \sum_{d=0}^{D} W_d f_{in}\left(\Lambda_d^{-1/2} A_d \Lambda_d^{-1/2} \otimes M_d\right), \tag{1}$$

where D represents the maximum distance of the graph, f_{in} and f_{out} represent the input and output values of the feature map, $\otimes$ represents the multiplication function, A and d mark the d-order adjacency matrix of the joint pair, and the result of the normalization operation is represented by A_d. W_d and M_d indicate adaptive adjustment parameters. It plays an important role in the realization of boundary adjustment and convolution operations. In order to extract temporal features, we insert a $L \times 1$ convolutional layer in the shallow layer to fuse the space information of the joint points between adjacent frames. In the process of temporal feature extraction, L represents the length of the time window, which is a predefined hyperparameter. Each time unit and space unit are followed by a BatchNorm module and a ReLU module to form a whole with this structure.

3.2. Residual Unit. Literature [38] proposed a structure called bottleneck, which cleverly uses the advantages of conv1×1 and is placed in the front and back positions of the common convolution part to reduce the number of feature channels in the convolution operation. In this paper, we cleverly used the bottleneck structure, abandoning the original time and space modules, and found in the experiment that the improved structure is significantly faster in model training and parameter calculation. For example, the input and output channels are 256, the channel reduction rate $r = 4$, and the time window size $L = 9$. Then, the total number of parameters involved in the calculation of the original structure is $256 \times 256 \times 9 = 589824$. If the bottleneck structure is adopted, the total number of parameters involved in the calculation is $256 \times 64 + 64 \times 64 \times 9 + 64 \times 256 = 69632$. Comparing the two, it can be seen that the bottleneck structure reduces the number of parameters calculated by the original structure by 8.5 times. Finally, we propose a new PartAtt block to enhance the generalization ability of the model. An example of a bottleneck structure frame is shown in Figure 3.

Considering that the time module and the space module in the original structure cannot integrate the features well, we connect the time and space modules with the residual structure to construct the ResGCN unit. The specific residual connection structure is shown in Figure 4. The Module residual module adopts a jump connection mode, the Block residual module adopts the mode of connecting before and after, the Dense residual module integrates the connection mode of the Module residual module and the Block residual module, making the structure more compact and saving calculation costs.

3.3. Multistream Input Structure. As we know from the bottleneck structure framework, each layer of input can be represented by a set of hyperparameters. In the first layer, we

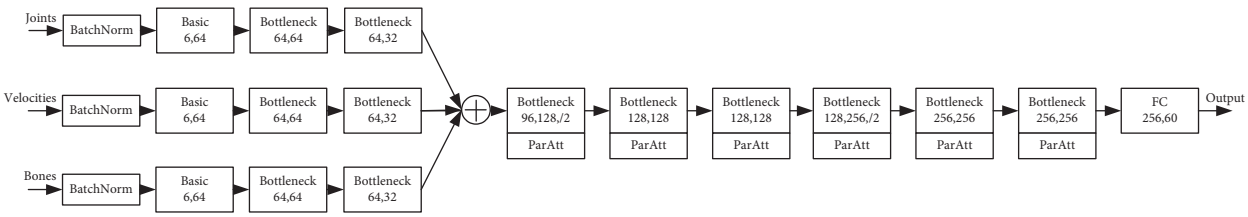
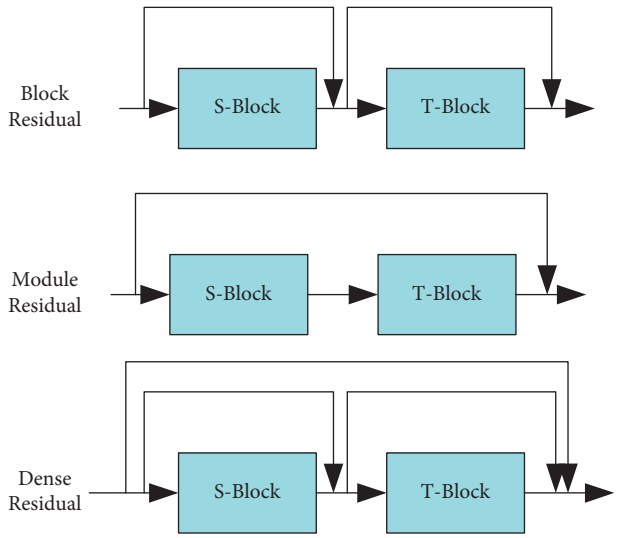


Figure 3: Yoga action recognition network fused with bottleneck structure.



S-Block: Spatial Convolution Block
T-Block: Temporal Convolution Block

Figure 4: Three types of residual structure.

usually use basic operations to process the original input data. The second layer starts to design the bottleneck structure to filter the output data of the previous layer, and the difference in the design of the bottleneck structure is the different number of channels between the input and output. The third and fourth layers also use the bottleneck structure, but the only difference is that each layer is followed by a PartAtt unit. By introducing the PartAtt unit, all the position information of the extracted feature vector is preserved. In the decoding process, the encoding can be performed directly by the PartAtt mechanism, which reduces the intermediate steps of traditional decoding and solves the problem of feature loss. Secondly, in the PartAtt mechanism, each step of encoding and decoding directly accesses the source feature library, which realizes the direct feature tradition of encoding and decoding and shortens the exchange in feature transfer. In addition, the time step is set to 2 in the input stage of the third and fourth layers to further reduce the complexity of parameter computation and prevent overfitting problems.

Furthermore, in high-precision models, input data generally require a multistream architecture for presentation. For example, the dual-stream input architecture mentioned in [39] incorporates both joint data and skeletal data as inputs, and decision selection is made after multiple streams of inputs. This approach is adopted by most researchers because it is effective in improving model performance. However, the multistream architecture does not control the computational cost well, and the large amount is data input, parameter exchange, and variable calculation in the multistream framework, which invariably increases the huge computational volume. Therefore, our action recognition model adopts a multistream architecture in the pretraining stage, with a total of three input branches, and each input branch feature is fused with mainstream features in a pass-through tandem manner. This structure not only preserves the skeleton features to a great extent, but also makes the model more concise in its vertical structure and easier to converge when the model is trained.

In the data preprocessing stage, we mainly used the methods proposed in [29, 40] for reference. In the motion recognition method based on bone joint points, data preprocessing is very critical. In our work, preprocessing mainly revolves around joint positions, motion speeds, and bone characteristics. Suppose that a video of the action sequence is collected. According to the action sequence, the spatial coordinate set is $X = \{x \in \mathbb{R}^{C\times T\times V}\}$, where C represents the coordinates, T represents the frame, and V represents the joints. You can also get the set of relative positions of bones in space $R = \{r_i | i = 1, 2, \ldots, V\}$, where $r_i = x[:,:, i] - x[:,:, c]$, x $[:,: c]$ represents human bones and spinal joints. Combining the sets R and X into one set can be input into the multistream branching framework as the joint positions in action recognition. In addition, two sets of speeds of each joint can be obtained $F = \{f_t | t = 1, 2, \ldots, T\}$ and $S = \{S_t | t = 1, 2, \ldots, T\}$, where $f_t = x[:, t + 2,:] - x[:, t,:]$ and $S_t = x[:, t + 1,:] - x[:,]$. Each motion feature of each joint can be represented by the two sets of feature vectors F and S, and this is input into the multistream branch frame as a motion stream. The basic characteristics of bones include length $L = \{L_i | i = 1, 2, \ldots, V\}$ and angle $A = \{A_i | i = 1, 2, \ldots, V\}$. The angle and length of the bone can be calculated through the bone displacement relationship $l_i = [:,:, i] - x[:,:, i_{adj}]$, where the first joint of i_{adj} represents the adjacent joint. The calculation equation for the angle obtained by conversion of the customs clearance equation is as follows:

$$a_{i,w} = arccos\left(\frac{l_i, w}{\sqrt{l_{i,x}^2 + l_{i,y}^2 + l_{i,z}^2}}\right), \tag{2}$$

where $w \in \{x, y, z\}$ represents space coordinates.

3.4. Partial Perception Framework. Long short-term memory neural network (LSTM) was proposed by Hochreiter [41] in 1997.

LSTM is a derivative of Recurrent Neural Network (RNN). Since 2010, it has been proven that RNN has been successfully applied to speech recognition [42], language modeling [43], and text generation [44]. However, the disappearance of gradients and explosions makes RNN difficult to apply to long-term dynamics research. As an improved network of RNN, LSTM can handle this problem well. LSTM gives the network a lot of freedom, so that the network memory unit has an adaptive solution to learn and update information, which greatly improves the performance of some perception networks.

Assume that $X = (x_1, x_2, ..., x_n)$ represents an input sentence composed of word representations of n words. In every position t, the RNN produces a hidden layer h in the middle denoted as y_t, and the hidden state h_t uses a nonlinear activation function to update the previously hidden layers h_{t-1} and the input x_t, as shown below:

$$\begin{aligned} y_t &= \sigma\left(W_y h_t + b_y\right), \\ h_t &= f\left(h_{t-1}, x_t\right), \end{aligned} \tag{3}$$

where W_y and b_y are the parameter matrices and vectors learned during the training process, and σ represents the elementwise softmax function.

The LSTM unit includes an input gate i_t, a forget gate f_t, an output gate o_t, and a memory unit c_t to update the hidden state h_t, as shown below:

$$\begin{aligned} i_t &= \sigma\left(W_i x_t + V_i h_{t-1} + b_i\right), \\ f_t &= \sigma\left(W_f x_t + V_f h_{t-1} + b_f\right), \\ o_t &= \sigma\left(W_o x_t + V_o h_{t-1} + b_o\right), \\ c_t &= f_t \odot c_{t-1} + i_t \odot \tanh\left(W_c x_t + V_c h_{t-1} + b_c\right), \\ h_t &= o_t \odot \tanh\left(c_t\right), \end{aligned} \tag{4}$$

where $\odot$ is a kind of function which is similar to the multiplication operation, V represents a matrix related to weight, and b represents the learning vector. To increase the model's performance, morpheme training was carried out on two LSTMs. The first one is a morpheme that begins on the left and works its way to the right; the next one is a reverse duplicate of a character. Before passing to the next layer, the outputs of the forward and reverse passes are combined in series. Finally, the prediction value is observed using the activation function.

After understanding the partial perception algorithm LSTM, it was inspiring, because in the human body recognition process, the human skeleton will be divided into multiple parts. Each part is an interconnected joint. These parts composed of joints are made by hand, for the graph convolution to be able to explore the relationship between these parts and extract the corresponding spatial features of the joint points. To obtain the information of a point in GCN, it is necessary to start from the field of that point. According to the adjacency matrix in the field, the skeleton data is automatically segmented, and then all the feedback information is input to the next joint point to complete the capture of the feature points of the entire human skeleton. Through this operation, the defects of manual design features are avoided, and the spatial features on the time series are obtained. [45].

If an ordinary convolutional neural network is used, all parts will be merged into a whole for feature extraction of convolution operations. Partial perception networks can divide joints into different departments and capture individual features for each part. Separately extracting features in this way helps to explore the connection between parts, that is, the spatial-temporal relationship between joints. The structure of our proposed spatial-temporal graph convolutional network-based yoga action recognition is shown in Figure 5.

4. Experiments

4.1. Data Collection. Before the establishment of the yoga posture database, we referred to yoga courses and training materials to find a reasonable grading system to assess the risk of yoga postures. As mentioned in [46, 47], the researcher compared the physical extensibility and commonality of action between the different postures. It was also approached in terms of breathing rate, posture intensity, and meditation. Also, we interviewed a yoga instructor who showed us all the standard yoga poses and broke down each pose. From his experience's we learned that currently there are 6 main yoga poses such as standing, forward bending, sitting, twisting, back bending, and supine. Each movement determines a different level of body stretch. In the study of this paper, the grading mainly revolves around these movements; our experimental scoring is based on the depth camera directly in front as the main interface. The specific grading is shown in Table 1.

In preparing the yoga dataset, we invited a professional instructor for standard yoga posture data collection. Then we invited participants who had one year of yoga experience and those who had no previous yoga experience to divide into two groups and complete each group of movements under the guidance of the instructor. In the process of data collection, not only the body posture but also the duration and number of movements of each yoga posture were collected. The yoga duration refers to the total time from when the breathing is adjusted until after the posture is completely relaxed. The number of movements refers to the sum of all the postures done during the training period, except for some correction of the postures by the instructor. The Azure Kinect DK was used to collect yoga movement data. The data is then manually calibrated by us after the data collection is completed and the data is split. In order to enhance the validity of the data, we added confidence parameters in the coding process. The data collection results are shown in Table 2.

4.2. Model Training. In the model training process, we used Pytorch to implement yoga movement recognition and

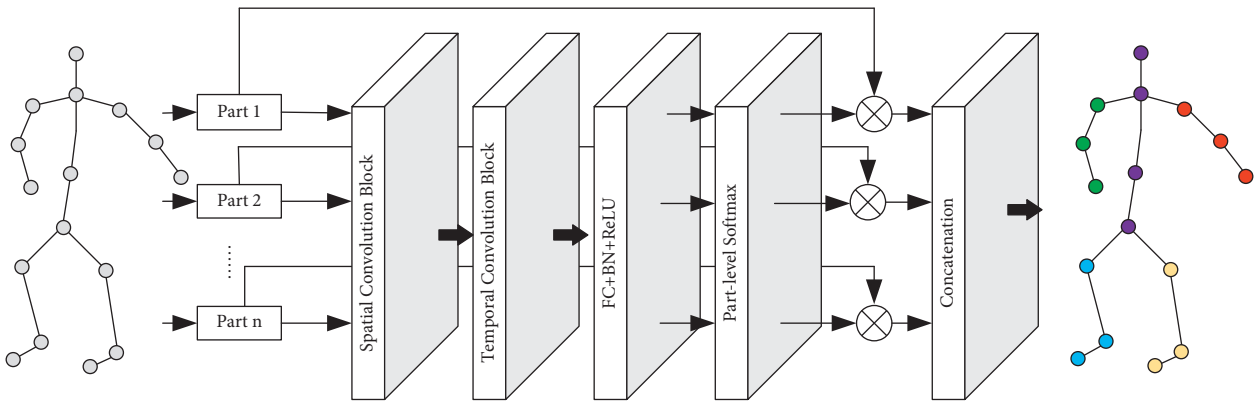


Figure 5: The overall network structure of yoga action recognition.

Table 1: Yoga action grading details.

Posture	Grade	Frontal	Sagittal
Standing	1	8	5
Sitting	2	6	6
Supine	3	10	9
Twisting	4	6	8
Forward bending	5	7	10
Back bending	6	9	6

Table 2: Yoga action time and frequency.

Posture	Experienced yogis		Inexperienced yogis	
	Ave time (s)	Ave frequency	Ave time (s)	Ave frequency
Standing	15	5	10	5
Sitting	31	6	16	6
Supine	23	8	14	8
Twisting	16	6	9	6
Forward bending	32	4	11	4
Back bending	27	5	12	5

grading. First, we used Openpose to extract the skeleton information from the yoga video dataset, and in each frame of the video we obtained the spatial coordinate information of each of the 14 joints. Then we use the heat map as the basis for pose estimation and perform secondary feature capture on the human skeleton. Then each frame of data is arranged in the temporal dimension to correlate the features between the joints from the temporal dimension. Finally, the skeletal joint features are fused using the average prediction score and the weights are estimated in a progressive ranking. We set different learning rates at different epochs. At the beginning of training, the learning rate is set to 0.05 to adapt to the training speed of the data. Then the learning rate is set to 0.01 at epoch = 30 to speed up the learning speed; after that, the learning rate is gradually reduced at epoch = 50 and epoch = 60 to find the optimal solution. The specific parameters in the model training are shown in Table 3. All the work is done in Ubuntu 16.04 and the whole training and prediction process is done with NVIDIA TITAN *X* GPU support on Intel Xeon E5-2620 CPU.

Table 3: Training parameter settings.

Parameter	Value
Epoch	20
Dropout rate	0.5
Initial learning rate	0.05
Learning rate (epoch = 30)	0.01
Learning rate (epoch = 50)	0.002
Learning rate (epoch = 60)	0.0004
Weight attenuation coefficient	0.0002
Momentum	0.9

4.3. Experimental Result. For the experimental data collection, we collected 50 experienced yogis and 50 inexperienced yogis. And the data was split according to the previous solution. The sensitivity, specificity, precision, and accuracy of skeletal features were captured in the data in the split starting from each frame. The experimental results are shown in Table 4. The standard yoga movements were decomposed on a larger scale, making it traceable in the validation set. Based on the above statistical results, higher

Table 4: Yoga action recognition results.

Posture		Sensitivity	Specificity	Precision	Accuracy
Standing	Experienced yogis	0.98	0.99	0.95	0.99
	Inexperienced yogis	0.71	0.91	0.91	0.98
Sitting	Experienced yogis	0.89	0.98	0.93	0.98
	Inexperienced yogis	0.66	0.91	0.91	0.98
Supine	Experienced yogis	0.94	0.99	0.89	0.99
	Inexperienced yogis	0.78	0.93	0.91	0.97
Twisting	Experienced yogis	0.96	0.98	0.91	0.99
	Inexperienced yogis	0.77	0.94	0.89	0.99
Forward bending	Experienced yogis	0.93	0.99	0.87	0.99
	Inexperienced yogis	0.69	0.91	0.89	0.98
Back bending	Experienced yogis	0.91	0.99	0.92	0.97
	Inexperienced yogis	0.72	0.92	0.86	0.97

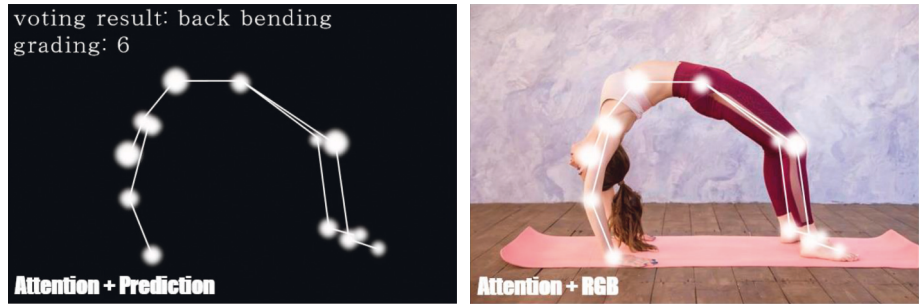


Figure 6: The effect of yoga action recognition and grading system.

Table 5: Yoga action grading results.

Posture	Grade	Experienced yogis (%)	Inexperienced yogis (%)
Standing	1	91	80
Sitting	2	93	77
Supine	3	89	75
Twisting	4	93	76
Forward bending	5	93	71
Back bending	6	92	72
Total ave	90	75	

sensitivity values represent more experience in yoga training and also predict closer standardization of yoga poses.

From the above experimental results, we can see that the recognition accuracy of all yoga poses is close to 1. And the accuracies, as a kind of random error, all keep above 0.86, which proves that the model performance is still great. The gap between experienced yogis and inexperienced yogis is mainly in sensitivity and specificity. Experienced yogis scored higher in both metrics, representing the more standardized yoga poses. The yogis are captured by the depth camera while practicing yoga. Real-time skeletal joint tracking is performed on the captured video. Finally, the yoga movements are recognized with the training model and then matched with the database to generate a grading score. The specific recognition effect is shown in Figure 6.

In addition, we also made corresponding statistics in the grading, as shown in Table 5.

Table 5 demonstrates that the average grading accuracy of experienced yogis in the whole set of yoga poses is higher than that of inexperienced yogis. The yoga posture with the greatest difference was forward bending, followed by back bending. Because of the difficulty of these two poses, it was difficult for inexperienced yogis to achieve the standard poses, so the accuracy of poses grading was lower. The above experimental results favorably prove the effectiveness of the grading system in this paper, which can give yogis feedback and remind them to change their postures if the yoga movements are not standard.

5. Conclusion

In this paper, we found that, with the popularity of the Internet, people's lifestyles have also changed, and many people choose to learn yoga by watching videos on the Internet. For yoga beginners, learning yoga online in this way without the direct guidance of an instructor, there is a high chance that the yoga poses will be substandard. Highly difficult yoga poses are likely to be disabling for beginners. To address this potential risk, we propose a yoga posture recognition and grading system based on spatial-temporal

graph convolutional neural network. We first use LSTM network to label yoga practice videos frame by frame. Then we extract the skeletal joint point features sequentially with graph convolution and then obtain the connection between joints from arranging each video frame in spatial-temporal dimension and correlating the joint points in each frame with neighboring frames for spatial-temporal information. Finally, through experiments, it is proved that our method can accurately identify yoga poses and grade them accordingly and can identify whether the yoga poses are standard or not and at the same time give feedback to yogis in a timely manner to prevent injuries to the body caused by nonstandard poses.

For deep learning algorithms, the larger the number of datasets, the better the accuracy of the model obtained from training. Since there is no specific dataset for yoga poses at present, the number of homemade datasets in this paper is small, which is the shortcoming of the work in this paper. Making datasets is a tedious and time-consuming task. In our future work, we will gradually increase the number of datasets, and at the same time, we will invest more efforts in the field of data preprocessing.

Data Availability

The dataset can be accessed upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This paper was supported by science and technology research project of Colleges and Universities in Inner Mongolia Autonomous Region, Project no. NJSY22477.

References

[1] Yogi Times, "Demographics & statistics of the yoga industry December 28," 2020, https://www.yogitimes.com/article/unstoppable-trend-yoga-infographic-business.

[2] P. Salmon, E. Lush, M. Jablonski, and S. E. Sephton, "Yoga and mindfulness: clinical aspects of an ancient mind/body practice," *Cognitive and Behavioral Practice*, vol. 16, no. 1, pp. 59–72, 2009.

[3] J. B. Webb, C. B. Rogers, and E. V. Thomas, "Realizing Yoga's all-access pass: a social justice critique of westernized yoga and inclusive embodiment," *Eating Disorders*, vol. 28, no. 4, pp. 349–375, 2020.

[4] Roy Morgan Research, "Yoga participants profile," 2016, https://www.roymorgan.com/findings/7004-yoga-is-the-fastest-growing-sport-or-fitness-activity-in-australia-june-2016-201610131055.

[5] Roy Morgan Research, "Yoga participation stretches beyond Pilates and aerobics March 30," 2018, https://www.roymorgan.com/findings/7544-yoga-pilates-participation-december-2017-201803290641.

[6] Ipsos Public Affairs, "2016 yoga in America study," 2017, https://www.yogaalliance.Org.

[7] J. K. Thompson, L. J. Heinberg, M. Altabe, and S. Tantleff-Dunn, "The scope of body image disturbance: the big picture [J]," *Exacting beauty: Theory, Assessment, and Treatment of Body Image Disturbance*, American Psychological Association, United States, pp. 19–50, 1999.

[8] A. Borden and C. Cook-Cottone, "Yoga and eating disorder prevention and treatment: a comprehensive review and meta-analysis," *Eating Disorders*, vol. 28, no. 4, pp. 400–437, 2020.

[9] E. A. Impett, J. J. Daubenmier, and A. L. Hirschman, "Minding the body: yoga, embodiment, and well-being," *Sexuality Research and Social Policy*, vol. 3, no. 4, pp. 39–48, 2006.

[10] L. Mahlo and M. Tiggemann, "Yoga and positive body image: a test of the Embodiment Model," *Body Image*, vol. 18, pp. 135–142, 2016.

[11] N. Piran and D. Neumark-Sztainer, "Yoga and the experience of embodiment: a discussion of possible links," *Eating Disorders*, vol. 28, no. 4, pp. 330–348, 2020.

[12] A. E. Cox and T. L. Tylka, "A conceptual model describing mechanisms for how yoga practice may support positive embodiment," *Eating Disorders*, vol. 28, no. 4, pp. 376–399, 2020.

[13] M. Singleton, *Yoga Body: The Origins of Modern Posture Practice*, Oxford University Press, United Kingdom, 2010.

[14] A. R. Jain, *Selling Yoga: From Counterculture to Pop Culture*, Oxford University Press, United Kingdom, 2015.

[15] J. B. Webb, E. R. Vinoski, J. Warren-Findlow, M. I. Burrell, and D. Y. Putz, "Downward dog becomes fit body, inc.: a content analysis of 40 years of female cover images of Yoga Journal," *Body Image*, vol. 22, pp. 129–135, 2017.

[16] N. Bhalla and D. Moscowitz, "Yoga and female objectification: commodity and exclusionary identity in U.S. Women's magazines," *Journal of Communication Inquiry*, vol. 44, no. 1, pp. 90–108, 2020.

[17] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas, "Discriminative density propagation for 3d human motion estimation[C]," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pp. 390–397, IEEE, San Diego, CA, USA, June 2005.

[18] R. Rosales and S. Sclaroff, "Learning body pose via specialized maps[C]," in *Proceedings of the Advances in Neural Information Processing Systems 14*, pp. 1263–1270, Vancouver, British Columbia, Canada, December 3-8, 2001.

[19] A. Agarwal and B. Triggs, "Monocular human motion capture with a mixture of regressors[C]," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)-Workshops*, p. 72, September 2005.

[20] L. Ren, G. Shakhnarovich, J. K. Hodgins, H. Pfister, and P. Viola, "Learning silhouette features for control of human motion," *ACM Transactions on Graphics*, vol. 24, no. 4, pp. 1303–1331, 2005.

[21] I. Cohen and H. Li, "Inference of human postures by classification of 3D human body shape[C]," in *Proceedings of the 2003 IEEE International SOI Conference. Proceedings (Cat. No. 03CH37443)*, pp. 74–81, IEEE, Nice, France, October 2003.

[22] R. Rosales, M. Siddiqui, J. Alon, and S. Sclaroff, "Estimating 3D body pose using uncalibrated cameras[C]," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, p. I, December 2001.

[23] K. Grauman, G. Shakhnarovich, and T. Darrell, "Inferring 3D structure with a statistical image-based shape model[C]," in

Proceedings of the Ninth IEEE International Conference on Computer Vision, p. 641, Nice, France, October 2003.

[24] X. Chen and M. Koskela, "Skeleton-based action recognition with extreme learning machines," *Neurocomputing*, vol. 149, pp. 387–396, 2015.

[25] M. Liu and J. Yuan, "Recognizing human actions as the evolution of pose estimation maps[C]," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1159–1168, Salt Lake City, UT, USA, June 2018.

[26] W. Du, Y. Wang, and Y. Qiao, "Rpan: an end-to-end recurrent pose-attention network for action recognition in videos[C]," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3725–3734, Venice, Italy, October 2017.

[27] J. Yu, M. Jeon, and W. Pedrycz, "Weighted feature trajectories and concatenated bag-of-features for action recognition," *Neurocomputing*, vol. 131, pp. 200–207, 2014.

[28] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data[C]," *Proceedings of the AAAI conference on artificial intelligence*, vol. 31, no. 1, pp. 4263–4270, 2017.

[29] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, "Skeleton-based action recognition with spatial reasoning and temporal stack learning[C]," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Springer, Cham, pp. 103–118, 2018.

[30] W. Zhu, C. Lan, J. Xing, and Y. Li, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks[C]," *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.

[31] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Proceedings of the European conference on computer vision*, Springer, Cham, pp. 816–833, 2016.

[32] C. Li, Q. Zhong, D. Xie, and S. Pu, "Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation[J]," arXiv preprint arXiv:1804.06055, 2018.

[33] U. Iqbal, M. Garbade, and J. Gall, "Pose for action-action for pose[C]," in *Proceedings of the 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pp. 438–445, IEEE, Washington, DC, USA, June 2017.

[34] D. C. Luvizon, D. Picard, and H. Tabia, "2d/3d pose estimation and action recognition using multitask deep learning [C]," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5137–5146, Salt Lake City, UT, USA, June 2018.

[35] B. Xu, J. Li, Y. Wong, Q. Zhao, and S. M. Kankanhalli, "Interact as you intend: intention-driven human-object interaction detection[J]," *IEEE Transactions on Multimedia*, vol. 22, no. 6, pp. 1423–1432, 2019.

[36] B. Wan, D. Zhou, Y. Liu, R. Li, and X. He, "Pose-aware multi-level feature network for human object interaction detection [C]," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9469–9478, Seoul, Korea (South), November 2019.

[37] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition [C]," in *Proceedings of the Thirty-second AAAI conference on artificial intelligence*, Louisiana, New Orleans, USA, February 2018.

[38] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition[C]," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[39] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12026–12035, Long Beach, CA, USA, June 2019.

[40] Y. F. Song, Z. Zhang, and L. Wang, "Richly activated graph convolutional network for action recognition with incomplete skeletons[C]," in *Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP)*, pp. 1–5, IEEE, Taipei, Taiwan, September 2019.

[41] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[42] O. Vinyals, S. V. Ravuri, and D. Povey, "Revisiting recurrent neural networks for robust ASR[C]," in *Proceedings of the 2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4085–4088, IEEE, Kyoto, Japan, March 2012.

[43] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, "Recurrent neural network based language model[C]," *Interspeech*, vol. 2, no. 3, pp. 1045–1048, 2010.

[44] I. Sutskever, J. Martens, and G. E. Hinton, "Generating text with recurrent neural networks[C]," in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, Washington, USA, June 2011.

[45] J. Liang and G. Zuo, "Taekwondo action recognition method based on partial perception structure graph convolution framework[J]," *Scientific Programming*, vol. 2022, pp. 1–10, 2022.

[46] U. H. Graneheim and B. Lundman, "Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness," *Nurse Education Today*, vol. 24, no. 2, pp. 105–112, 2004.

[47] M. Bengtsson, "How to plan and perform a qualitative study using content analysis," *NursingPlus Open*, vol. 2, pp. 8–14, 2016.