HSiPu^2 - A New Human Physical Fitness Action Dataset for Recognition and 3D Reconstruction Evaluation

Chuanlei Zhang[1] Lixin Liu[1] Minda Yao[1] Wei Chen[2, 3*] Dufeng Chen[4]
Yuliang Wu[5]
[1]Tianjin University of Science and Technology.
[2]China University of Mining and Technology (Beijing).
[3]China University of Mining and Technology.
[4]Beijing Geotechnical and Investigation Engineering Insititute.
[5]Sichuan Staff University of Science and Technology.
{a17647, lixinliu0327, yaominda1995}@gmail.com
{chenwdavior, dfchen_bjiei, Luckfan2002}@163.com

Abstract

In this paper a human physical fitness action feature dataset named HSiPu^2 is introduced, which contains 8,044 action data sequences and 80,440 images. The dataset is built for three physical fitness actions, which are sit-up, push-up and pull-up, and each action data has two categories corresponding to standard and non-standard actions. Two cameras work as sensors to capture features from different views. HSiPu^2 data set facilitates the evaluation of the performance of machine learning algorithms used in recognition and evaluation problems related to human behaviours recognition. The dataset is freely and publicly available online. A new recognition method based on deep learning techniques, including the two-branch multi-stage convolutional neural networks (CNNs) and long short-term memory (LSTM) networks with attention, are employed to learn the long-term dependencies from videos for human physical fitness action recognition on the HSiPu^2. For comparison purpose, traditional machine learning models, including Decision Tree, Random Forest, SVM, Bagging, GBDT, AdaBoost, XGBoost and Voting, are utilized for the action recolonization. Furhtermore, HSiPu^2 also can serve as a dataset to evaluating a human action 3D model.

1. Introduction

Recognition of human actions in videos has been popular in the research of computer vision community due to its important significance in theoretical study and great value in practical application. Especially, the recognition of human motions and gestures is attracting increasing emphasis in the area of computer vision [1]. Human behavior recognition has achieved good results in smart home, abnormal behavior detection, gait recognition and so on. However, there are few researches on physical pose recognition, and there is no relevant dataset. In the school or the army, physical training and assessment on the physical exercise poses of students or soldiers are usually necessary. However, it takes enormous human resources to manage and evaluate these human actions in an artificial manner, and for some reasons, factitious assessment is insufficient in objectivity. Therefore, researches on recognition of the motion poses have obtained more and more attention recently. Motivated by these facts, a human physical fitness action feature dataset named HSiPu^2 is introduced. The dataset with ground truth labels is built for three human physical fitness actions, which are sit-up, push-up and pull-up, and each action data has two categories corresponding to standard and non-standard actions. Two cameras work as sensors to capture features from different views. Further, in this paper, a physical fitness action recognition method from video is proposed, which are based on two-branch multi-stage CNN [2] and LSTM-Attention architecture [3]. At the same time, the dataset could be utilized to compute a 3D model of human human action, which can facilitate the analysis of human actions and interactions.

In summary, the contribution of our work is listed as follows:

(1) A new human physical fitness action dataset, named HSiPu^2, with ground truth labels is built. The human physi-

*Correspondence author: Wei Chen (chenwdavior@163.com)

cal fitness action dataset is released publicly to ensure future research in the area.

(2) HSiPu^2 is a multi-view dataset, where two cameras are adopted as sensors to capture human physical fitness action classification features in different views.

(3) Deep learning techniques, including two-branch multi-stage CNN and long short-term memory (LSTM) network with attention mechanism are adopted, to learn the long-term dependencies from video frames for human physical fitness action recognition.

(4) Extensive experiments have been conducted and the results validated the effectiveness of the proposed method.

(5) Two classic 3D reconstruction models are evaluated on the HSiPu^2.

2. Related Work

Prior to the emergence of the methods of deep learning, most of the traditional methods of human behaviour recognition include three main tasks. The first one is feature extraction of human behaviours, and spare spatial-temporal interest points are extracted. The second step is to characterize the behaviours. Thirdly, the features are classified by pattern classifiers like Decision Tree, Support Vector Machine (SVM), K Nearest Neighbors(K-NN), and so on [4]. In recent years, researchers have been making attempts to apply the modes of deep neural networks, such as Convolutional Neural Network (CNN), to video behaviour recognition. Vida *et al*. [5] investigated the challenging problem of action recognition in videos and proposed a new component-based method for analysing the video content. Convolutional reexisting neural networks (ConvRNNs) enables robust spatial-temporal information processing for the recognition of contextual video, but the computation cost is high with an inefficient training. Jung *et al*. [6] came up with "adaptive detrending" (AD) for temporal normalization in order to accelerate the training of ConvRNNs, especially of convolutional gated reexisting unit (ConvGRU). Dense trajectories, as the most commonly used local feature descriptor of video, has a superior performance in action recognition for a variety of datasets. However, its computation is complicated and the algorithm requires much space for storage, thus finally constraining its application. You *et al*. [7] optimized the algorithm of action recognition based on the optimized features of the dense trajectories. Athira [8] proposed a signer independent novel vision-based gesture recognition system which is capable of recognizing single-handed static or dynamic gestures, double-handed static gestures, and finger spelling words of Indian Sign Language (ISL) from the live video. In addition, an improved method for co-articulation elimination in finger-spelling alphabets was also put forward. Zhang *et al*. [4] put forward a Siamese neural network named as Motion-patch-based Siamese Convolutional Neural Network which is the first attempt to alleviate the influence of poor samples generated by the conventional methods of data augmentation and to enhance the motion information of the videos through data input in HAR systems. Mukherjee *et al*. [9] presented a new framework for the recognition of mid-air finger writing using web-cam video as input. A termination criterion based on the velocity of the fingertip was adopted as a delimiter to signal the completion of the air-writing gesture. Xu [10] proposed a learning method for the video semantic features. It integrated topological sparse coding of the image with the algorithm of dynamic time warping to improve the gesture recognition in videos. Itano *et al*. [11] presented a recognition framework for human actions from the video scenes from multiple viewpoints of the camera. Mliki *et al*. [12] introduced a new approach to recognizing the human activity from the UAV-captured video sequences. Hsueh *et al*. [13] employed deep learning techniques, including convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, to build up deep networks to expound the long-term dependencies for human behaviour recognition in a multi-view framework from the videos. He *et al*. [14] presented an action recognition method to heighten the feature representation of video frames by capturing local temporal context, and introduced chained residual pooling to perform multistage fusion of neighboring frames. Wu *et al*. [15] proposed a multi-teacher knowledge distillation framework to compress this model for action recognition of the compressed video. Most approaches to action recognition based on the video create the video-level representation by temporally pooling the features extracted at every frame. The pooling methods usually completely or partially neglect the dynamic information contained in the temporal domain, thus undermining the recognition capability of the resulting video representation. To figure out this drawback and probe into the importance of incorporating the temporal order information, Wang *et al*. [16] came up with a novel temporal pooling approach to aggregating the frame-level features, and employed the temporal convolution operation for the frame-level representations to extract the dynamic information. Amosov *et al*. [17] used the ensemble of deep neural networks to design a classifier for intellectual situations, managing to achieve detection and recognition of normal and abnormal situations in a continuous video stream of the security system. Jing *et al*. [18] put forward an efficient and straightforward approach, video you only look once (Video YOLO), to capture the overall temporal dynamics for action recognition in a single process from the whole video. The method is extremely time-saving. VideoYOLO-32 is able to process 36 videos per second, which is 10 times and 7 times faster, respectively than the prior 2D-CNN (Two-stream [19]) and 3D-CNN (C3D [20]) based models. Cao *et al*. [2] presented an approach to efficiently detecting the 2D

poses of multiple people in an image. A nonparametric representation called Part Affinity Fields (PAFs) was adopted to learn to associate body parts with individuals in the image. The nonparametric representation of the association between the key points can encode both the position and the orientation of human limbs. Furthermore, to better understand human action, sometimes, we need the information more than the major joints of the body. We have to know the full 3D surface of the body, hands and even the face. Thus, a lot research effort is paid to appropriate 3D models and training dataset. There are some researches on extracting such a human action 3D model from a single human action image [21].

3. HSiPu^2: A Dataset for Human Physical Fitness Actions

The dataset, named HSiPu^2 [3], for training, testing and verification in the paper are captured by our research team. The physical fitness actions in the experiment contain three types, which are push-up, pull-up and sit-up. For each action, two cameras are used for video capturing and one camera is in front and another one is on side. For each action type, both standard actions and non-standards actions are captured. The steps to obtain the data are as follows:

Step 1 Record all the actions into video files with two cameras. The captured video data is divided into 6 categories, which are sit-up front, sit-up side, push-up front, push-up side, pull-up front, pull-up side. The sample pictures of the dataset are demonstrated in Figure 1;

Step 2 Select one frame from every 6 frames of the video, that is, extract about 5 frames of images per second;

Step 3 Cut out the main part of the human behaviour in the picture.

After data collection, the captured data are processed to facilitate the feature extraction of human key points. The steps to process the data are as follows:

The captured images are manually annotated, and the sequence of pictures marked as the same action is placed in the same folder. Table 1 demonstrates the amount of each action data. Each type data of action is annotated, and they are divided them into standard action data and non-standard action data. The model of the two-branch multi-stage CNN is used to extract 17 key point features of each image into one-dimensional vector and store them in a file. The extracted information includes human key points identification, relative coordinates (x coordinate and y coordinate), and confidence, so that 68 elements of information are stored into one-dimension vector, which can be extracted from one single image; 10 frames are taken as a sequence to create the dataset, that is, the human key points data of every 10 frames are used as the input of the model. Finally, the amount of each type of data sequence is demonstrated in Table 2. Totally there are 8044 cropped data sequences. In order to ensure future research in the area, the dataset is publicly released at: https://github.com/mindayao/HSiPu2.

(a) push-up side. (b) pull-up front. (c) sit-up side.

(d) push-up front. (e) pull-up side. (f) sit-up front.

Figure 1. The sample pictures of the dataset HSiPu^2.

	Push-up	Pull-up	Sit-up
Front	443	723	1564
Side	2786	1692	2317

Table 1. The amount of each action in dataset HSiPu^2.

4. Two-Branch Multi-Stage CNN and LSTM-Attention Based Human Action Recognition

The architecture of the proposed two-branch multi-stage CNN and LSTM-Attention based human action recognition method is demonstrated in Figure 2. For spatial dimension, the two-branch multi-stage CNN is utilized to extract the human key points features. The image is first input to VGG-19, generating a set of feature maps F. Feature maps F words as input to the two-branch multi-stage CNN. For temporal dimension, the sequence of key point features works as input to obtain the temporal features. And the LSTM-Attention mode is employed to recognize the physical fitness action categories. There are about three major modules, and they are discussed in detail as follows.

4.1. Spatial Feature Extraction through Two-Branch Multi-Stage CNN

The conventional method of human action estimation is top-down, which refers to detecting the human body area

	Push-up front	Push-up side	Pull-up front	Pull-up side	Sit-up front	Sit-up side
Standard	72	606	139	331	290	411
Non-standard	279	1896	450	1073	827	1670

Table 2. The amount of data sequence in dataset HSiPu^2.

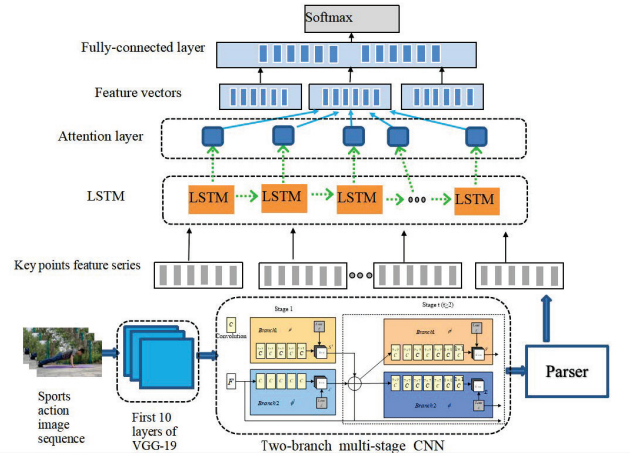


Figure 2. The proposed human physical fitness recognition architecture.

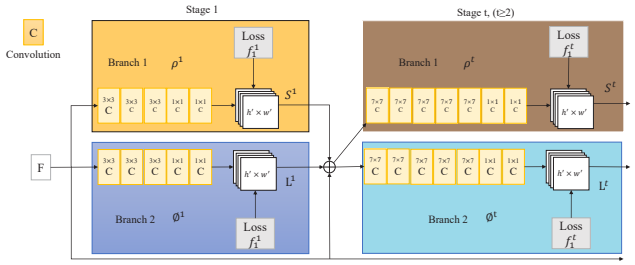


Figure 3. The architecture of two-branch multi-stage CNN [2].

first, and then detecting the key points of the human body in the area. As it is necessary to perform forward key point detection for each detected human body area, the speed is quite slow. Cao *et al*. [2] presents the Realtime multi-person 2D pose estimation method, which is a first bottom-up representation of association scores via Part Affinity Fields (PAFs), a set of 2D vector fields that encode the orientation and position of limbs on the image domain. Based on the detected human key points and Part Affinity Fields, these human key points can be mapped to different individuals using the greedy inference algorithm. The network structure is demonstrated in Figure 3.

There are two branches in the network which are top branch and bottom branch. The top branch is employed to predict the confidence maps, and the bottom branch is utilized to predict the affinity fields. At beginning, the image is input to VGG-19, generating a set of feature maps F. Then the feature maps F works as input to the first stage of each branch. At the first stage, the network generates a set of detection confidence maps $S^1 = \rho^1(F)$ and a set of part affinity fields $L^1 = \phi^1(F)$. ρ^1 and ϕ^1 are the CNNs for inference at first stage. In each following stage, the predictions from both branches in the last stage (as shown in Equation 1 and Equation 2), along with the original image features F, are concatenated and employed to create refined predictions,

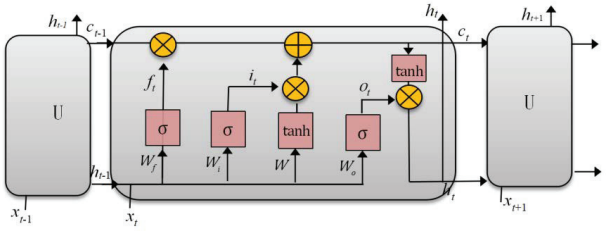


Figure 4. The proposed human physical fitness recognition architecture.

$$S^t = \rho^t(F, S^{t-1}, L^{t-1}), \forall t \geqslant 2 \quad (1)$$

$$L^t = \phi^t(F, S^{t-1}, L^{t-1}), \forall t \geqslant 2 \quad (2)$$

4.2. LSTM Neural Network

LSTM is a kind of special RNN model [22][23][24]. LSTM projects the input to the hidden state and the hidden state to the output, so as to effectively learn the dynamic information of the input human body key point feature sequence. For RNN, it is sensitive to short-term inputs as the hidden layer has only one state h. Instead, there are three control gates in LSTM to learn long-term dependent information. The state c in LSTM store the long-term state, which solves the problem of RNN with only one hidden layer state h. Figure 4 is the LSTM unit structure. There are forgetting gate f_t, input gate i_t, output gate o_t and one memory unit.

In Figure 4, the yellow nodes indicate operating the vectors one by one. "$\bigotimes$" means multiplication, calculating the dot product between two vectors and addition "$\bigoplus$" means to calculate the sum. The pink node indicates activation operation. They are $tanh$ function and σ function. x_t works as the input data at moment t, h_t works as the output state value of the LSTM unit at moment t, c_t works as the candidate value of the memory unit at moment t, i_t works as

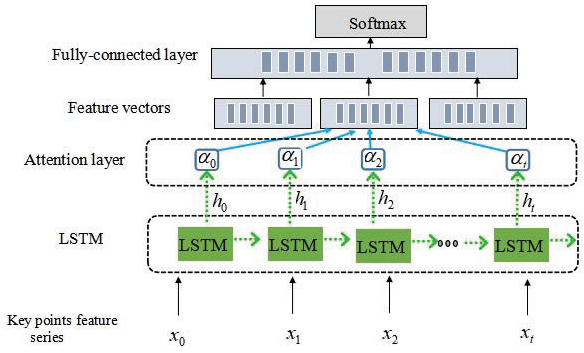


Figure 5. The proposed human physical fitness recognition architecture.

the state value of input gate at moment t, f_t works as the state value of forgetting gate at moment t, W works as the corresponding weight, b works as the corresponding bias parameter. The state value of the memory unit is regulated by the input gate and the forgetting gate.

4.3. Attention Mechanism

In the field of natural language processing and image processing, attention mechanism is widely used [24][25][26]. Several attention mechanisms have been put up with by scholars, and the classification effect is remarkably improved. The introduction of attention mechanism to LSTM makes it possible to discover the internal relationship among the feature information. It can generate the classification result by weighted average, which can improve the recognition performance of the model. For a series of weight parameters, the major idea of attention mechanism is to learn the importance of each element from the sequence, and pay different attention to the elements based on their importance. The attention mechanism can greatly improve the classification performance of the model. Furthermore, the it can also be used to investigate how the information in the input sequence affects the final output results.

Therefore, when designing a model, a layer of attention network following LSTM is used to extract temporal features. A LSTM-Attention classification model can be demonstrated in Figure 5.

5. Experiment and Result Analysis

To evaluate the effectiveness of our proposed physical fitness action recognition method based on multi-stage convolutional neural networks (CNNs) and LSTM-Attention model, the architecture is trained and tested on the dataset HSiPu^2. In this section, the implementation details of experiment are described and the experiment results are analyzed.

In our experiments, the dataset HSiPu^2 are randomly divided into the verification set and training set according to the ratio of 1:9. The key point information extracted is relative coordinates, that is, the coordinates relative to the length and width in the image. In order to improve the generalization ability of the model, the sequence images are cut randomly during the process of extracting human key points to ensure that the relative positions of human key points change in turn, so as to extend the data set.

In this section, we also evaluated two classic 3D reconstruction models on the dataset HSiPu^2. The two models are SMPL-X and PIFuHD.

5.1. Experiment Configuration Details

The experiments are conducted with the following implement details. First, the data are read from the TFRecord files and are put into the memory buffer. The batch size is set to 128. This training group is put into the LSTM-Attention model to perform training operations. Then the loss is calculated and the optimizer back propagation is used to reduce the loss and adjust the network parameters of each layer. The optimizer is the Adam optimizer provided by TensorFlow. The initial value of learning rate is set to 1e-4, and the optimization the model is learning rate exponential decay.

5.2. Comparation Experiment Result and Analysis

In order to verify the effectiveness of the proposed LSTM-Attention method, the model is compared it with several traditional machine learning models, including Decision Tree [27][28], Random Forest [29], SVM [30], Bagging [31], GBDT [32][33], AdaBoost [34], XGBoost [35] and Voting.

Decision Tree. Decision tree model is often used to solve classification and regression problems. Decision tree is a tree structure, which can be binary tree or non-binary tree. Each non-leaf node represents a test on a feature attribute, each branch represents the output of the feature attribute in a certain range, and each leaf node stores a category. The process of using decision tree to make decision is to start from the root node, test the corresponding characteristic attributes in the item to be classified, and select the output branch according to its value until it reaches the leaf node, and take the category stored by the leaf node as the decision result. For decision trees, data processing is often simple or unnecessary and it is able to work with both data-type and regular-type attributes. In addition, it is insensitive to missing values and it can process irrelevant feature data. The decision tree only needs to be built once and used repeatedly, and the maximum calculation time of each prediction does not exceed the depth of the decision tree. However, it also has many disadvantages. It is difficult to predict the contin-

uous fields. For sequence data, there is a lot of processing work to do. In addition, when there are many categories, the accuracy of prediction will decrease.

Random Forest. Random Forest is to establish a forest in a random way. There are many decision trees in the forest, and there is no correlation between each decision tree in the random forest. After the forest is obtained, when a new input sample is entered, each decision tree in the forest shall be judged separately to see which category the sample belongs to, and then to see which category is selected the most, so as to predict which category the sample belongs to. Random Forest can handle data with high dimensions without feature selection and it is simple to implement. When creating the random forest, the generalization error is used with unbiased estimation, and the model has strong generation ability. During the training process, interaction between the features can be detected. For unbalanced datasets, it can balance the error. Even if there is a lot of features are missing, the accuracy of the prediction results can be maintained. However, the random forests have been shown to be overfitted in some classification and regression problems. For the data of attributes with different values, attributes with more value division will have a greater impact om the random forest, so the attributes weights produced by the random forest on such data are not credible.

Support Vectors Machine (SVM). SVM is a discriminant classifier defined by the classification hyperplane. In other words, given a set of labeled training samples, the algorithm will output an optimal hyperplane to classify the new test samples. Nonlinear mapping is the theoretical basis of SVM method. SVM makes use of inner product kernel function instead of nonlinear mapping to high dimensional space. The optimal hyperplane for feature space division is the goal of SVM, and the idea of maximizing the classification margin is the core of SVM method. Support vector is the training result of SVM. It is the support vector that plays a decisive role in SVM classification decision making. SVM is a novel small-sample learning method with a solid theoretical foundation. It basically does not involve probability measure and law of large numbers, so it is different from the existing statistical methods. In essence, it avoids the traditional process from induction to deduction, realizes the efficient "transduction reasoning" from training samples to prediction samples, and greatly simplifies the usual problems such as classification and regression. However, SVM algorithm is difficult to implement for large-scale training samples. Because SVM uses quadratic programming to solve support vector, and solving quadratic programming will involve the calculation of m-order matrix (m is the number of samples), when the number of M is very large, the storage and calculation of this matrix will consume a lot of machine memory and operation time. In addition, it is difficult to solve multi-classification problems with SVM. The classical support vector machine algorithm only gives the two-class classification algorithm, but in the practical application of data mining, the multi-class classification problem is generally solved.

Bagging. Bagging is an integrated algorithm to improve classification by combining randomly generated training sets. Bagging only uses a subset of the training set as the current training set for each training data (random sampling is put back). Each training sample can appear for many times or not in a training set. After T times of training, T different classifiers can be obtained. When a test sample is classified, T classifiers are called respectively to obtain T classification results. Finally, the class with high occurrence frequency in T classification results is assigned to the test sample. This sampling method is called bootstrap, which is to use the limited sample data to reconstruct a new sample sufficient to represent the original sample distribution through repeated sampling. The advantage of Bagging is that when there is noise data in the original sample, one-third of the noise samples will not be trained through Bagging sampling. Bagging is useful for noise-affected classifiers. Therefore, Bagging can reduce the variance of the model and is not easily affected by noise. It is widely used in unstable models or models prone to overfitting.

Gradient Boost Decision Tree (GBDT). GBDT is an additive model based on boosting ensemble. In training, the forward distribution algorithm is used for greedy learning. Each iteration learns a cart tree to fit the residual between the prediction result of the previous T-1 tree and the real value of the training sample. GBDT has the flexibility to handle all types of data without the need for feature normalization. Compared with SVM, the prediction accuracy is relatively high in the case of relatively few parameters. In addition, compared with the traditional decision tree, the residual calculation of each step is equivalent to learning along the optimal direction, which increases the weight of the error instance and uses fewer features to make decisions to prevent overfitting. However, it is a serial process, which cannot be parallelized, and does not fit the linear model.

AdaBoost. AdaBoost makes good use of the weak classifier for cascade and it can use different classification algorithms as weak classifiers. In addition, the accuracy of the prediction results is high. Compared with Bagging algorithm and Random Forest algorithm, AdaBoost fully considers the weight of each classifier. However, the number of AdaBoost iterations, which is the number of weak classifiers, is not easy to set, and can be determined by cross-validation. Data imbalance leads to a decrease in classification accuracy. Training is time-consuming, it is best to reselect the current classifier each time the segmentation point.

XGBoost. The basic idea of XGBoost is the same as GBDT, but some optimizations have been done, such as default missing value processing, adding second derivative infor-

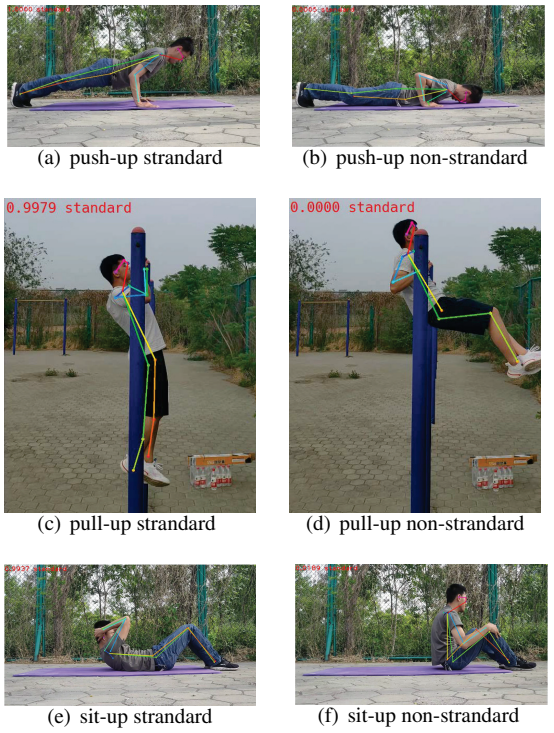
mation, regular term, column sampling, and parallel computing. When looking for the best segmentation point, considering that the traditional greedy method of enumerating all possible segmentation points of each feature is too inefficient, XGBoost implements an approximate algorithm. The general idea is to enumerate several candidates that may become the segmentation point according to the percentile method, and then find the best segmentation point from the candidates according to the above formula for finding the segmentation point. XGBoost takes account into the sparse value of the training data, and can specify the default direction of the branch for the missing value or the specified value, which can greatly improve the efficiency of the algorithm. The feature columns are sorted and stored in the memory in the form of blocks, which can be reused in iteration; although the boosting algorithm iteration must be serial, it can be parallelized when processing each feature column. Storing in the feature column method can optimize the search for the best segmentation point, but when calculating gradient data in rows, it will cause discontinuous access to the memory. In severe cases, it will cause cache miss and reduce the efficiency of the algorithm. XGBoost can first collect data into the internal buffer of the thread, and then calculate it to improve the efficiency of the algorithm. XGBoost also considers how to effectively use disks when the amount of data irrelatively large and the memory is insufficient. It mainly combines multithreading, data compression, and fragmentation methods to improve the efficiency of the algorithm as much as possible.

Voting. The weak classifier used in voting is Logistic Regression, Decision Tree and Gaussian naive bayes. The final decision-making method is hard voting classifier.

The comparison experiment result is demonstrated in the Table 3. It can be found that the LSTM-Attention model is much better than the traditional machine learning models. The reason lies in two aspects. One is that Long Short-Term Memory (LSTM) network, due to its power in modelling the dynamics and dependencies in sequential human action key points feature data, has superior performance in human action recognition. Another one is the introduction of attention mechanism. Usually, different human action key points feature data sequences often have different informative human action key points, and in the same sequence, the informativeness degree of a key point may also vary over the frames. Therefore, it is a good idea to selectively focus on the informative human action key points in each frame, and try to ignore the features of the irrelevant key points, since the irrelevant key points do not contribute to physical fitness action recognition, and even bring in much noise that can deteriorate the performance of action recognition method.

5.3. Live Demo of the Proposed Recognition Method

Thanks to the high efficiency of two branch multi-stage CNN feature extraction, a live demo system is constructed based on the proposed human physical fitness action recognition method to show that our approach can be used in practice in real-time. The system is tested with the captured video. The results are demonstrated in the following pictures. The upper left corner of the picture shows the probability that this motion pose is standard. If the probability is greater than 0.5, it instructs that the motion pose is standard. The sample results are demonstrated in Figure 6.



(a) push-up strandard (b) push-up non-strandard

(c) pull-up strandard (d) pull-up non-strandard

(e) sit-up strandard (f) sit-up non-strandard

Figure 6. The sample pictures of the dataset HSiPu^2.

5.4. 3D Reconstrction Effect with the Dataset

SMPL-X [21] and PIFuHD [36] are two models which can reconstruct human action 3D from a single image. SMPL-X, that extends SMPL with fully articulated hands and an expressive face, is a new, unified, 3D model of the human body. PIFuHD is Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. It is a multi-level framework that performs joint reasoning over holistic information and local details to arrive at high-resolution 3D reconstructions of clothed humans from a single image without any additional post processing or side information. PIFuHD achieves this by incrementally propagating global context through a scale pyramid as an implicit 3D embedding. This avoids making premature de-

	Push-up front	Push-up side	Pull-up front	Pull-up side	Sit-up front	Sit-up side
Decision Tree	84.06%	80.55%	73.75%	77.96%	79.90%	80.35%
Random Forest	93.62%	83.33%	80.14%	89.83%	85.64%	86.60%
SVM	73.70%	66.66%	70.91%	79.66%	81.81%	74.10%
Bagging	90.84%	86.11%	78.01%	81.36%	79.90%	91.07%
GBDT	94.82%	94.44%	85.11%	89.83%	89.47%	85.71%
AdaBoost	95.22%	88.89%	84.39%	91.53%	86.60%	87.50%
XGBoost	94.92%	94.38%	88.39%	92.32%	92.60%	88.50%
Voting	85.26%	83.33%	70.92%	77.97%	77.51%	83.93%
Proposed method	**98.80%**	**97.22%**	**95.74%**	**96.61%**	**97.23%**	**97.31%**

Table 3. The accuracy of models.

cisions about explicit geometry that has limited prior approaches.

From the Figure 7 and 8, we can find that HSiPu^2 can serve as dataset from which we can construct 3D of human physical actions, and SMPL-X perform better to reconstruct the three actions, except the sit-up front view.

Figure 7. 3D reconstruction of human activity from the dataset with PIFuHD.

Figure 8. 3D reconstruction of human activity from the dataset with SMPL-X.

6. Conclusion

A new human physical fitness action dataset, named HSiPu^2, with ground truth labels is built and the data have two categories corresponding to standard and non-standard human physical fitness actions. The data has been publicly released to ensure future research in the area. Further, in this paper, a human physical fitness action recognition method based on two-branch multi-stage CNN and LSTM-Attention network structure is proposed. The proposed method considers the temporal concepts to learn the spectral-spatial information so that human behavior can be recognized precisely. A lot of experiments are conducted to verify the proposal, comparing with other traditional machine learning models. The result proves that the recognition accuracy and loss of this method have good performance and the result outperforms the traditional models. Moreover, two models, SMPL-X and PIFuHD, are evaluated on the dataset to reconstruct human action 3D from a single image. In future, we will further improve the performance of the method for video data in complex environment. The sequence of dataset created has 680 features. We will investigate how to do the feature selection and fusion to improve the recognition rate. Moreover, we will extend our system by further investigating physical fitness action assessment for multiple people.

Acknowledgments. This work is supported by the key project of Tianjin natural Science Foundation [grant numbers 18JCZDJC32100] and Tianjin Science and Technology Commissioner project [grant numbers 19JCTPJC51100].This work is also funded by National Natural Science Foundation of China [grant number 51874300], National Natural Science Foundation of China and Shanxi Provincial People's Government Jointly Funded Project of China for Coal Base and Low Carbon [grant number U1510115], and the Open Research Fund of Key Laboratory of Wireless Sensor Network and Communication, Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, [grant numbers 20190902 and 20190913].

References

[1] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C Kot. Global context-aware attention lstm networks for 3d action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1647–1656, 2017. 1

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 1, 2, 4

[3] Chuanlei Zhang, Lixin Liu, Qihuai Xiang, Jianrong Li, and Xuefei Ren. Sports pose estimation based on lstm and attention mechanism. In *International Conference on 5G for Future Wireless Networks*, pages 538–550. Springer, 2020. 1, 3

[4] Yujia Zhang, Lai Man Po, Mengyang Liu, Yasar Abbas Ur Rehman, Weifeng Ou, and Yuzhi Zhao. Data-level information enhancement: Motion-patch-based siamese convolutional neural networks for human activity recognition in videos. *Expert Systems with Applications*, 147:113203, 2020. 2

[5] Vida Adeli, Ehsan Fazl-Ersi, and Ahad Harati. A component-based video content representation for action recognition. *Image and Vision Computing*, 90:103805, 2019. 2

[6] Minju Jung, Haanvid Lee, and Jun Tani. Adaptive detrending to accelerate convolutional gated recurrent unit training for contextual video recognition. *Neural Networks*, 105:356–370, 2018. 2

[7] Wenwan You, Junqi Guo, Ke Shan, and Yazhu Dai. A novel trajectory-vlad based action recognition algorithm for video analysis. *Procedia Computer Science*, 147:165–171, 2019. 2

[8] PK Athira, CJ Sruthi, and A Lijiya. A signer independent sign language recognition with co-articulation elimination from live videos: an indian scenario. *Journal of King Saud University-Computer and Information Sciences*, 2019. 2

[9] Sohom Mukherjee, Sk Arif Ahmed, Debi Prosad Dogra, Samarjit Kar, and Partha Pratim Roy. Fingertip detection and tracking for recognition of air-writing in videos. *Expert Systems with Applications*, 136:217–229, 2019. 2

[10] Shuping Xu, Lixin Liang, and Chengbin Ji. Gesture recognition for human–machine interaction in table tennis video based on deep semantic understanding. *Signal Processing: Image Communication*, 81:115688, 2020. 2

[11] Fernando Itano, Ricardo Pires, Miguel Angelo de Abreu de Sousa, and Emilio Del-Moral-Hernandez. Human actions recognition in video scenes from multiple camera viewpoints. *Cognitive Systems Research*, 56:223–232, 2019. 2

[12] Hazar Mliki, Fatma Bouhlel, and Mohamed Hammami. Human activity recognition from uav-captured video sequences. *Pattern Recognition*, 100:107140, 2020. 2

[13] Yu-Ling Hsueh, Wen-Nung Lie, and Guan-You Guo. Human behavior recognition from multiview videos. *Information Sciences*, 517:275–296, 2020. 2

[14] Feixiang He, Fayao Liu, Rui Yao, and Guosheng Lin. Local fusion networks with chained residual pooling for video action recognition. *Image and Vision Computing*, 81:34–41, 2019. 2

[15] Meng-Chieh Wu and Ching-Te Chiu. Multi-teacher knowledge distillation for compressed video action recognition based on deep learning. *Journal of Systems Architecture*, 103:101695, 2020. 2

[16] Peng Wang, Lingqiao Liu, Chunhua Shen, and Heng Tao Shen. Order-aware convolutional pooling for video based action recognition. *Pattern Recognition*, 91:357–365, 2019. 2

[17] OS Amosov, SG Amosova, YS Ivanov, and SV Zhiganov. Using the ensemble of deep neural networks for normal and abnormal situations detection and recognition in the continuous video stream of the security system. *Procedia Computer Science*, 150:532–539, 2019. 2

[18] Longlong Jing, Xiaodong Yang, and Yingli Tian. Video you only look once: Overall temporal convolutions for action recognition. *Journal of Visual Communication and Image Representation*, 52:58–65, 2018. 2

[19] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*, 2014. 2

[20] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2

[21] G. Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10967–10977, 2019. 3, 7

[22] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2016. 4

[23] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *arXiv preprint arXiv:1409.3215*, 2014. 4

[24] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 4, 5

[25] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014. 5

[26] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 5

[27] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984. 5

[28] John Ross Quinlan. Probabilistic decision trees. In *Machine Learning*, pages 140–152. Elsevier, 1990. 5

[29] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 5

[30] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 5

[31] Alaa Tharwat, Tarek Gaber, Yasser M Awad, Nilanjan Dey, and Aboul Ella Hassanien. Plants identification using feature fusion technique and bagging classifier. In *The 1st International Conference on Advanced Intelligent System and Informatics (AISI2015), November 28-30, 2015, Beni Suef, Egypt*, pages 461–471. Springer, 2016. 5

[32] Jerry Ye, Jyh-Herng Chow, Jiang Chen, and Zhaohui Zheng. Stochastic gradient boosted distributed decision trees. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 2061–2064, 2009. 5

[33] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems*, 30:3146–3154, 2017. 5

[34] Trevor Hastie, Saharon Rosset, Ji Zhu, and Hui Zou. Multi-class adaboost. *Statistics and its Interface*, 2(3):349–360, 2009. 5

[35] T. Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. 5

[36] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93, 2020. 7