ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
https://doi.org/10.17559/TV-20240708001834
Original scientific paper

Human Action Recognition Using Explainable Features and Sparse Motion History Images

Wei YANG, Yitong ZHOU, Jianying XIONG, Shiwei ZHANG, Lei ZHANG, Leiyue YAO*

Abstract: This study proposes a novel approach to human action recognition (HAR) via depth sensor data. We introduce explainable features derived from skeleton sequences and a sparse motion history image (SMHI) to effectively capture motion characteristics. Our method addresses the limitations of current approaches by reducing the computational requirements while maintaining high accuracy. We propose a SlowFast-like network that combines these features for efficient HAR. Experiments on three datasets demonstrate the effectiveness of our approach, which achieves competitive accuracy with fewer features. The method also ensures user privacy by relying solely on skeleton data. This research contributes to the theoretical advancement of HAR and its practical application in various fields.

Keywords: data augmentation; human action recognition, motion feature matrix; skeleton-based HAR, video encoding

1 INTRODUCTION

Vision-based applications of human action recognition (HAR) have made significant progress because of its widespread applications in human-machine interaction, security monitoring, surveillance systems, etc. [1, 2]. Especially since the emergence of deep learning technology, HAR research and its applications have increased in popularity in recent years.

Due to the excellent feature-extracting ability of the deep neural network (DNN), the most immediate work is replacing the pre-popular classifier with a convolutional neural network (CNN) or long short-term memory (LSTM) for higher accuracy. These simple classifier-replacing works achieved better performances in experiments and applications than the pre-popular classifier. Another simple but effective solution is putting all frames of an action sample into a DNN to train the model or obtain the predicted result. This solution is called the "end-to-end" method, and it was once popular because of its simplicity in algorithm and application.

In the pre-deep-learning era, the common HAR method was "extracting the features of human action quantifying these features using the quantified features to train a support vector machine (SVM) predicting or classifying a new human action by the well-trained SVM" [3]. The kernel of this type of method involves extracting the features of a human action. Many well-designed hand-crafted features were proposed, some of which continue playing important roles in the current pattern recognition field. These handcrafted features can be divided into two categories: global features and local features. The extraction of global features uses the top-down methodology, treats the human body as an entity and typically extracts features via the interframe difference method or background subtraction method. The typical features are the motion energy image (MEI) [4], motion history image (MHI) [4], static history image (SHI) [5], and motion history volume (MHV) [6]. The most obvious merit of a global feature is its meaningful semantic information, which can be used to explain the entire model or improve its accuracy. However, global features heavily depend on the hardware and are extremely sensitive to illumination, occlusion, and dynamic background. Thus, HAR methods that use only global features are always dataset-oriented and not robust. To address the issues of global features, local features that are mathematically calculated are proposed. Representative features are spatio-temporal interest points (STIPs) [7], histograms of oriented gradients (HOGs) [8], and 3D Harris [9]. Compared with global features, local features are much less sensitive to illumination, occlusion, and dynamic background [2]. Thus, local features are more suitable for real-world applications. However, local features cannot reflect the high-level semantic information of human actions, which results in difficulties in method improvement. Local features are also typically computationally intensive, and their algorithms always have high complexity. This drawback partly prevents its application in real-time situations. In particular, mobile applications lack powerful computational ability.

As mentioned, in the deep learning era, due to the excellent feature extraction capabilities of CNNs and LSTM, many DNN-based methods and their variations [10-13] significantly outperform SVM-based methods. Depending on the powerful computational capacity of the graphics processing unit (GPU) and massive training samples, DNN-based methods can always achieve high recognition accuracy without complex algorithms. Unlike rul-based methods, DNN-based methods are driven by data and referred to as "end-to-end" methods. However, "end-to-end" methods soon face their bottleneck. An unexplainable characteristic of the deep feature is that DNN-based methods have great difficulties in fine-tuning. Although researchers have proposed many solutions, such as transformer learning [14-[16], few-shot learning [17-[19], and zero-shot learning [20][21], the most effective solution continues to be enlarging the training samples. Many researchers have paid special attention to data augmentation methods. In image classification, many data augmentation methods have been shown to be effective and widely applied in various applications, such as rotation, cutting angles, and disturbances [22]. However, data augmentation methods of video recognition are far more complex because of the temporal features and context information in a video. Thus, in most video recognition applications, pre-processing is essential. In this step, meaningful features are well designed, extracted, quantified, and stored in a matrix to fit the input requirements of the DNNs. For example, optical flow (OF) is a classic pre-processed image to describe motion

pixels in a video. Many early HAR methods used OF and its variations to realize action video classification [23,[24].

With the development and popularization of depth cameras, it has become much easier for researchers to extract human skeletons from videos. Since a human action can be considered the motion of the key joints and rigid bones [26], increasing amount of skeleton-based HAR methods have emerged. Y. Du et al. [27] converted human actions to colour images and classified these images via a CNN. The position of a joint in 3D coordinates (x, y, z) was innovatively stored in the three channels (R, G, B) of a colour pixel. In [28], the action image in [27] was improved by a tree structure skeleton image (TSSI). The TSSI can better preserve the spatial relationships of the joints, which helps improve the accuracy. In our previous works [29], [30], these methods were further improved, and a self-defined data structure named the Dense Joint Motion Matrix (DJMM) was used to store joint motion features with high precision. Since the advent of the GCN, various GCN-based methods [33-[35] have emerged to effectively model the human skeleton and improve the HAR accuracy. The advantages of using skeleton motion information to recognize human actions can generally be summarized in 3 points.

1) Compared with motion pixels, the human skeleton is strongly related to the acting subject. Using skeleton motion information can dramatically reduce the side effects of noisy inputs.

2) Using skeleton motion information can ignore the differences caused by the height, body shape, and other individual differences. All humans have similar skeleton structures.

3) Using the skeleton motion information, it is easy to mimic an action conducted by different people with different body sizes; i.e., data augmentation becomes easier.

Among numerous skeleton-based HAR methods, researchers proposed various motion features [36] and designed multifarious deep learning models [35], [37-39] to classify videos or recognize human actions. Although the proposed feature and deep learning model partly promote HAR to a new level, the universality of the application and the balance between efficiency and effectiveness remain challenging. In this work, we continue our previous research and prove that adopting the key motion feature is sufficient to recognize human actions. Several new motion features with explainable characteristics were introduced, many experiments were conducted to find the best combination of the new motion features and classic features, and a multi-scale CNN with high universality was constructed to recognize human actions. Our work has three major contributions:

1) A 3D float matrix was proposed and constructed using only features that have explainable physical or kinematic meanings and can be easily calculated from the coordinates of 25 joints.

2) An efficient multi-scale CNN and a series of data augmentation strategies were designed to prevent the model from overfitting and recognize human actions with different temporal durations.

3) Full experiments were conducted to find the best feature combination that can balance efficiency and effectiveness.

The remainder of this paper is structured as follows: Section 2 covers related works on the skeleton-based HAR method. Section 3 discusses the proposed features and an efficient multi-scale CNN. Section 4 evaluates the proposed method using three typical datasets. Finally, Section 5 presents the conclusions.

2 RELATED WORK

As mentioned, MEI, MHI, SHI and their variations or combinations are typical global features that are widely used for HAR and achieve good performance. However, a significant drawback prevents these global features from being applied in real scenes: when the motion trajectories overlap, the temporal or spatial information of the action may be lost. Thus, local features such as the 3D coordinates [27], [28], displacements [29], and motion directions [30] of each joint of the skeleton were introduced to solve the problem caused by trajectory overlaps. Many researchers have proposed features to classify actions or distinguish similar actions. For example, Liao L. C. et al. [31] used the angles of adjacent bones as a new local feature for HAR. Du Y. et al. [27] rearranged the joints and stored their coordinates in a sequence to represent the physiological information of the skeleton. Dhiman C. et al. [32] synthetically used 5 classical local features to improve the accuracy.

In the current data-driven era, if more motion features are used, the HAR model will be more accurate. Thus, many methods attempt to use as many features as possible, whereas few researchers are concerned about which feature combination is the most cost-effective and how to make further improvements based on the existing features. Here, we continue our previous work and propose 3 instructive rules for skeleton-based HAR methods.

1) Explainable features, especially those defined by classic physical or kinematic theories, are recommended for input matrix construction. Thus, the method can be fine-tuned in the input stage.

2) Regardless of the number of training samples, a data augmentation method based on explainable features is necessary because of the high cost of video sample labelling.

3) A multi-scale neural network with attention mechanisms helps improve the flexibility and accuracy of the HAR method, and because of the pre-processing work in the feature extraction stage, a deeper network is not better.

3 PROPOSED APPROACH

The proposed method follows the 3 aforementioned rules. As shown in Fig. 1, the work can generally be concluded in 3 steps. First, an action is quantified frame by frame and encoded as a 3D float matrix. Then, a data augmentation method generates new action samples with different temporal scales. Finally, a multi-scale CNN is proposed to recognize the action. As shown in Fig. 1a, the motion matrix and motion image are input into two networks, and the prediction confidences of these two networks of all action classes are averaged. The final result is the action with the highest average confidence score.

The key to our proposed method is how to extract explainable features and quantify and encode an action video as a 3D float matrix. It is the prior work of the data augmentation method and multi-scale CNN construction. This kernel work is also strongly related to the ablation experiments, which are conducted to determine the best combination of the selected features. The detailed introduction is described in the following section.

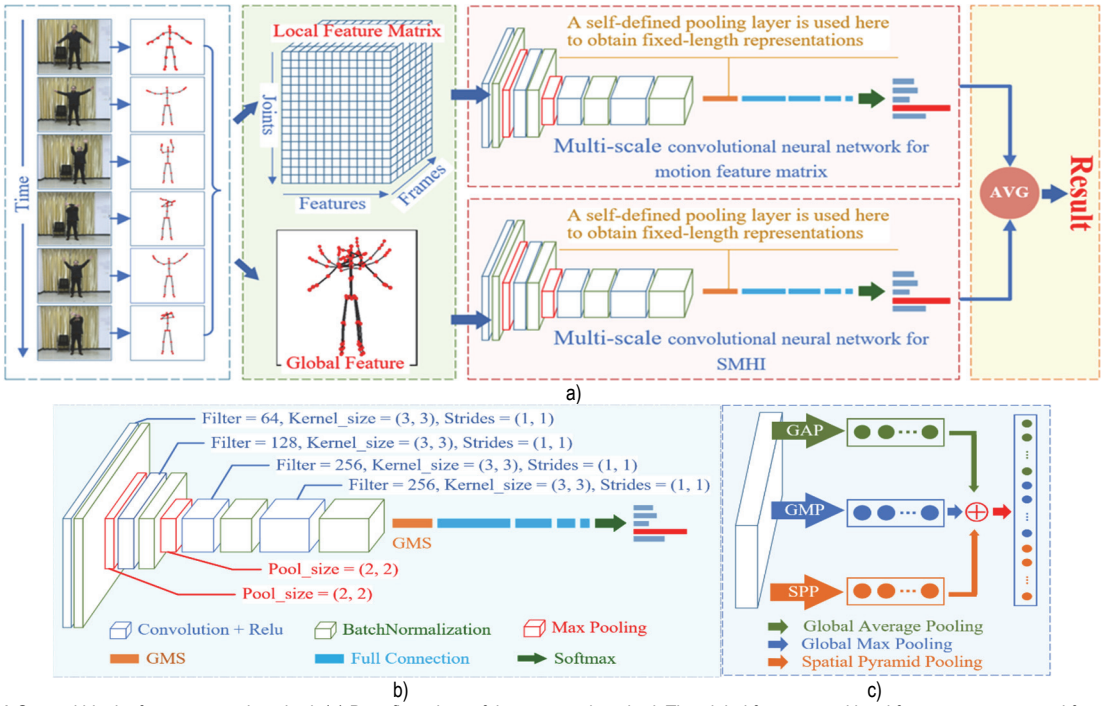


Figure 1 General block of our proposed method. (a) Data flow chart of the proposed method. The global features and local features are extracted from the skeleton sequence. Then, the global features and local features are put into two neural networks to recognize human actions. (b) Hyper parameters of the proposed CNN. The flatten layer of the classical CNN is replaced by a self-designed layer, called the GMS layer, which can satisfy the multi-scale learning purpose. (c) Structure of the self-defined GMS layer. The GMS layer is composed of a global average pooling (GAP) layer, a global max pooling (GAP) layer and a spatial pyramid pooling (SPP) layer

3.1 Feature Selection

In many prior methods, the 3D coordinates of joints are directly used as input parameters [26], [27]. As shown in Fig. 2, 25 joints in total can be captured by Kinect v2.0.

However, our previous works [29], [30] demonstrate that the related coordinate is the better choice because of its flexibility and universality. In the proposed method, it is recommended to use the coordinates of the "Spine Mid" joint in the first frame as the original point.

Figure 2 Diagram of 25 joints of the human skeleton captured by Kinect v2.0

Since the deep feature is unexplainable, the greatest challenge of "end-to-end" methods is manually fine-tuning the deep network instead of using algorithms or strategies. To address this issue, in the action encoding stage, we select explainable features with significant physical or kinematic meanings. These features include 5 quantified motion features and a global motion image.

1) Motion direction of a joint between two frames (θ). The physical or kinematic meaning of this feature is to record the direction variation of a joint in the duration of a frame. It contains both spatial and temporal information. The context information is essential for describing an action.

According to the translatability of the vector, the motion direction in the duration of frames $(j-i)$ can be projected on planes xy, xz and zy, and their values can be calculated via the law of cosines.

$$\theta^{i,j}=\begin{cases}\theta_{xy}^{i,j}=\dfrac{x}{\sqrt{\Delta x^2+\Delta y^2}}\\ \theta_{xz}^{i,j}=\dfrac{xz}{\sqrt{\Delta x^2+\Delta z^2}}\\ \theta_{zy}^{i,j}=\dfrac{z}{\sqrt{\Delta z^2+\Delta y^2}}\end{cases} \tag{1}$$

where $\{x,y,z\}$ is the position of a joint in the ith frame, and $\{\Delta x,\Delta y,\Delta z\}$ is the displacement of the joint on the x, y and z-axes among frames i and j.

2) Displacement of a joint between two frames (D). This feature records the spatial variation of a joint in a time period.

$$D=\sqrt{\left(x_i-x_j\right)^2+\left(y_i-y_j\right)^2+\left(z_i-z_j\right)^2} \quad (2)$$

where $\{x_i, y_i, z_i\}$ and $\{x_j, y_j, z_j\}$ are the coordinates of a joint in the i th and j th frames, respectively.

3) Motion speed of a joint between every two frames (D'). This feature indicates the motion intensity of a joint in a time period. From a physical perspective, it can be used to discriminate similar actions. Based on D, D' can be easily calculated via Equation 3.

$$D'=\frac{D}{0.3333\times n} \quad (3)$$

where 0.3333 is the duration time of one frame (Kinect v2 records video at 30 fps, i.e., there are 30 frames in a second), and n is the number of interval frames.

4) Joint angles in a frame (α). In the geometrical view, the set of joint angles of a frame can be partly considered a description of the human pose, whereas an MHI-like global feature can be quantified by storing the angles in a float array in a time sequence.

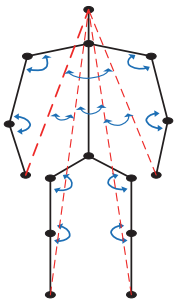
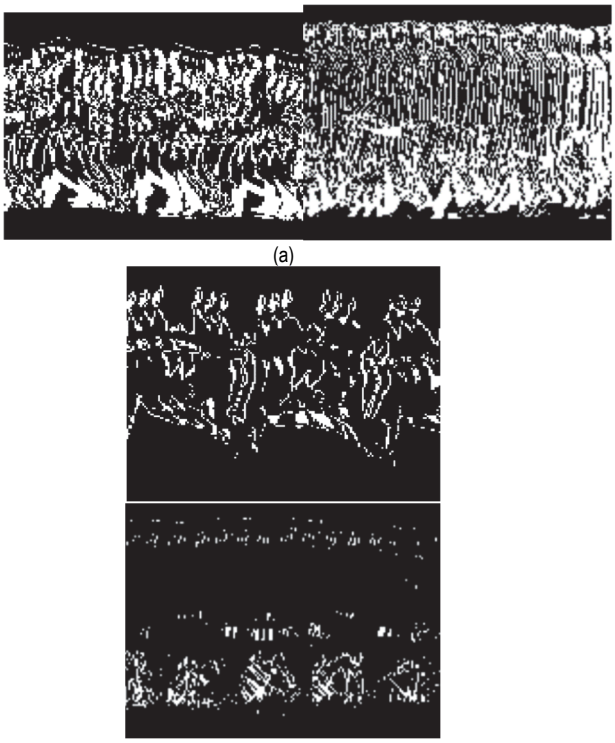


Figure 3 Joint angles in the proposed method

As shown in Fig. 3, there are two types of joint angles: one type consists of 3 adjacent joints, and the other consists of head, hand, and foot joints. The angles constituted by the red dotted lines are designed to help this feature better discriminate similar actions.

5) Variations of the proposed features. To recognize a human action, it is necessary to consider the immediate value of a feature in each frame and its final value in the last frame. The collection of the immediate value contains valuable context information, whereas the final value is the final status of a joint and may include qualitative spatial and temporal information. Thus, the immediate motion direction $\left(\theta^{i,i+1}\right)$, final motion direction $\left(\theta^{0,\tau}\right)$, immediate displacement $\left(D^{i,i+1}\right)$, and total displacement $\left(D^{0,\tau}\right)$ must be extracted for HAR.

6) Sparse motion history image (SMHI). Human actions, especially complex ones with context information, are difficult to recognize when only local features are used. Global features can well describe context information. The MHI, which is the most classic motion image, can be complementary for HAR. Since the temporal or spatial overlap problem is not negligible for global features, in the proposed method, the dense sampling method of MHI is abandoned; only the frames at every 10% interval of the action duration are used, and the motion image created in this manner is named the sparse motion histogram image (SMHI). Fig. 4b shows its comparative advantage over the typical MHI. Fig. 4a indicates that the MHIs may sometimes confuse an action with another similar action because of its dense sampling method. Thus, although this sparse sampling method is not the best solution for generating MHIs, it is sufficient for describing the motion tendency of an action and works as an effective complementary feature of the above 5 local features.



(a)

(b)

Figure 4 Comparison of MHI and SMHI. (a) MHIs of "run" and "walk". (b) SMHIs of "run" and "walk"

3.2 Data Augmentation and Motion Matrix Design

Using CNNs and pre-processing video data to recognize human action is a mainstream method of HAR. Compared to RNN-, LSTM- and GCN-based methods, its greatest advantage is the flexible data structure. Moreover, since our selected features are explainable, the action samples can be mathematically generated. In the proposed method, the data augmentation strategies are designed with 4 strategies:

1) Scaling the actor's skeleton to mimic an action that is conducted by people of different sizes. It is the simplest but most effective method to generate action samples. Because the proposed method is based on the human skeleton and people of different sizes have similar skeleton proportions, a new action sample can be easily generated by a) multiplying the 3D coordinates of each joint in each frame by a coefficient and b) calculating the quantified features and generating the SMHI, as introduced in Section 3.1.

2) Modifying an action's duration to mimic an action that is conducted by the same people at different speeds. When the CNN is used to recognize human actions, an important step is to standardize the size of all samples.

However, maintaining the duration of the actions on the same scale is very difficult, and the existing uniform methods, which are widely used in image classification, such as rotation, cutting and scaling, are not suitable for HAR. To retain the maximum motion information of an action, we propose an algorithm based on the bilinear interpolation theory to unify the temporal sizes of the action samples. Fig. 5 shows its flow chart, where TN is the uniform number of action durations, and TS is the frames of the current processing action sample.

3) Rotating an action's motion direction to mimic an action that is conducted by the same people at the same speed in different directions. Since the gravity line is always the *y*-axis of the depth image, we selected the "SpineMid" joint as the origin of the coordinates, and the coordinates of other joints in a depth frame are rotated at a certain angle around the y-axis. Then, all features of Section 3.1 can be similarly extracted. Eq. (4) is used to calculate the rotated coordinates of a joint.

$$\begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix} \cdot \begin{bmatrix} \mathrm{x} \\ \mathrm{y} \\ \mathrm{z} \\ 1 \end{bmatrix} \tag{4}$$

4) Generating more action samples using a combination of the above 3 strategies.

3.3 A Multi-Scale Remoulding Strategy for CNNs

To address the 3D float matrix with variable size, especially variables in the time dimension, we proposed a multi-scale CNN remoulding strategy using a spatial pyramid pooling (SPP) layer in our previous work. In this work, we improve the strategy by combining the SPP layer, GAP layer and GMP layer into the GMS layer. Compared with the previous work, the accuracy is improved because the GAP layer can effectively preserve the global features of the hidden layers, whereas the GMP layer can reserve the distinguishing features.

Fig. 6 shows the multi-scale structure of the GMS layer. The final output size of the GMS layer can be calculated using Eq. 5:

$$L = \sum_{i}^{n} N_f \times P_i \tag{5}$$

where L is the output length of the GMS layer, Nf is the number of feature maps, and $P = (P_0, P_1, \ldots, P_i)$ is the pooling layers in the SPP layer. Taking $P = [(4,4), (3, 3), (2, 2)]$ as an example, if we assume that the number of feature maps before the SPP layer is 128, the output size of the GMS layer is $(4\times4 + 3\times3 + 2\times2) \times 128 + 128 + 128 = 3968$.

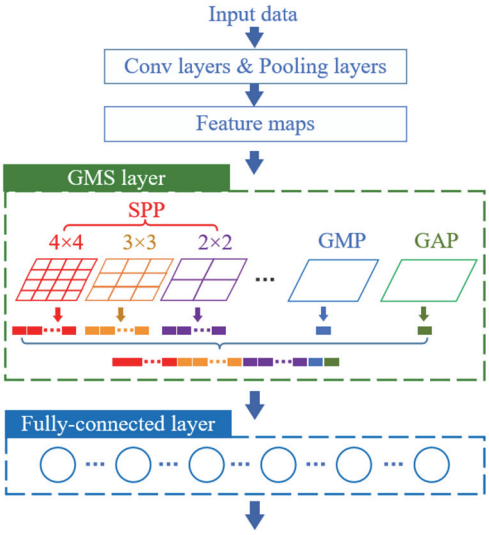


Figure 6 Structure of the GMS layer with 1 SPP layer, 1 GMP layer and 1 GAP layer. The pooling parameter of the SPP layer was set as [4, 3, 2]. Thus, the output of the GMS layer has a fixed length of 31 dimensions.

4 EXPERIMENTS AND EVALUATION

The method was implemented using the TensorFlow 2.3 GPU version and Keras. The experiments were conducted on a desktop computer with an Nvidia GTX-3060 GPU, Intel Core i7-10700K 3.70 GHz processor and 64 GB of RAM clocked at 3200 MHz.

To investigate the effectiveness of our proposed methods, we conducted experiments on three datasets: the Florence 3D actions dataset [40], UTKinect-Action 3D dataset [41] and HanYue Action 3D dataset (hereafter referred to as Florence-3D, UT-3D and HanYue-3D, respectively). The first two are small datasets, and the third one is a self-collected dataset. Eighty percent of the samples for each action in the dataset were used as training samples, and 20% of the samples for each action in the dataset were used as testing samples.

4.1 Datasets

Florence 3D Action Dataset: The dataset was collected at the University of Florence in 2012 and captured using a Kinect camera. It includes nine actions: wave, drink from a bottle, answer phone, clap, tight lace, sit down, stand up, read watch and bow. During the action data collection process, 10 subjects were asked to perform the above actions 2 - 3 times each, which resulted in 215 activity samples [40].

UTKinect Action Dataset: The dataset was collected using a single stationary Kinect sensor. There were 10 actions: walk, sit down, stand up, pick up, carry, throw, push, pull, wave hands and clap hands. There were 10 subjects, and each subject performed each action twice, which resulted in 199 activity samples [41].

HanYue Action 3D dataset: The dataset was collected using a Kinect v2.0 camera. It includes 15 simple action types: make a phone call, drink, wave hands, look at a watch, pat the dust off the clothe, fall, push a chair, jump in place, stand up, stand still, stand clap, walk, sit, sit still, and sit clap. Nine subjects were asked to perform the 15 activities 3 - 4 times. All positions of 25 joints in 3D coordinates provided by the Kinect v2.0 sensor were

recorded. There are 413 samples in total, and each action type contains 35 ~ 37 samples. Additionally, for action temporal localization purposes, 4 complex behaviour types were collected, each of which contained several of the above simple actions. The 4 complex behaviours are: "walk →fall→stand up→pat the dust off the clothes"; "sit still→ stand up→look at a watch→stand still→look at a watch→ wave hands"; "push a chair→drink→push a chair→drink →sit down→sit still→stand up→walk"; "sit clap→stand up →stand clap→jump→wave hands".

Table 1 Comparison of the test accuracy of different CNNs that were trained on the original samples and enlarged samples [29]

Dataset	Score Item	leNet-5	AlexNet	ZfNet	DenseNet121	VGG16	VGG19	ResNet50
Florence-3D	based on original samples	79.07%	74.42%	62.79%	20.93%	76.74%	58.14%	37.21%
	based on generated samples	83.47%	90.80%	92.50%	92.87%	87.95%	83.37%	88.80%
	Accuracy improved	4.40%	16.38%	29.71%	71.94%	11.21%	25.23%	51.59%
UT-3D	based on original samples	67.50%	67.50%	72.50%	50.00%	65.00%	37.50%	47.50%
	based on generated samples	82.36%	85.87%	91.84%	88.71%	85.44%	83.09%	94.87%
	Accuracy improved	14.86%	18.37%	19.34%	38.71%	20.44%	45.59%	47.37%

4.2 Evaluation of the Coordinate Transformation Strategy

To avoid interference caused by the installation positions of different cameras, we proposed two coordinate transformation methods: global-trans (G-trans) and local-trans (L-trans).

G-trans converts the coordinate system origin to the "SpineMid" joint in the first frame of an action. Then, all of the other frames were set according to the transformed origin of the coordinates to reset their coordinates. We suggest that G-trans should be used in calculating motion features, such as the motion direction of a joint between two frames, the motion speed of a joint between every two frames, and the motion speed of a joint between every two frames.

L-trans converts the coordinate system origin to the "SpineMid" joint in the current frame of an action. Since L-trans cannot represent the motion information of an action, we suggest that L-trans is better used to calculate static or geometrical features, such as the joint angles in a frame.

4.3 Evaluation of Data Augmentation Strategies

The small sample problem is a large obstacle to the application of deep learning technology. The 3 datasets that we adopted have few samples. UT-3D includes 119 action samples in total, and each action group contains only 20 samples. Florence-3D includes 215 action samples in total, and each action group contains only approximately 24 samples. MSR-3D consists of 20 action groups and 567 action samples in total, and each action group contains only 27 ~ 30 samples. The samples are insufficient to effectively train a CNN, and deep neural networks will suffer from overfitting.

In our previous works [29], [30], the duo-quadratic interpolation algorithm, a human skeleton scaling method and a temporal translation strategy were adopted to generate 3 types of samples. Tab. 1 shows the experimental results of our previous work, and the generated "creditable" samples are helpful for training CNNs. The term "credible" refers to samples that are both reasonable and valuable. Unlike traditional data augmentation methods, the samples generated using the original version can simultaneously mimic people of different sizes performing the same action, a certain action being conducted at different speeds, or both situations.

In this paper, we insist on generating "creditable" samples and improving the data augmentation strategies. As introduced in Section 3.2, rotation of the motion direction of an action is added as a new data augmentation strategy.

All of the following experiments were conducted on the amplified datasets.

4.4 Evaluation of the Proposed Multi-Scale CNN

To enable a CNN to learn deep features from multi-scale samples, we raised the GSM layer to replace the "Flatten" layer of the classical CNNs. Tab. 1 records the accuracies of 6 classical CNNs and their modified versions, which were tested on UT-3D and Florence-3D, respectively.

Tab. 2 implies that the remoulded CNN performed as well as or slightly better than its original version in most situations. Furthermore, the remoulded CNNs generally performed better when the samples were amplified at different scales. The oldest CNN-LeNet achieved competitive performance during the experiments because the input feature matrix was small. It is unnecessary to use a too-deep network. Under the guidance of this ideology, we proposed a shallow network (S-GMS) and fine-tuned it until it achieved satisfactory accuracy. The structure of S-GMS and its parameters are shown in Fig. 1b. In the remaining experiments, S-GMS was selected as the backbone.

Table 2 Comparison of the remoulded CNNs and original CNNs

	UT-3D	Florence-3D
LeNet-5	84.62%	77.50%
	82.05%	77.50%
AlexNet	84.62%	82.50%
	87.18%	85.00%
ZfNet	79.49%	70.00%
	82.05%	70.00%
VGG16	77.50%	85.00%
	80.00%	87.50%
VGG19	71.79%	85.00%
	61.54%	87.50%
S-GMS (Our proposed CNN)	87.18%	87.50%
	89.74%	90.00%

4.5 Ablation Experiment of Motion Feature Combinations

With the development of vision-based HAR methods, increasingly many features have emerged. Some researchers have adopted as many features as possible to improve the detection accuracy [32]. However, this is not the best solution because of the ever-increasing computing power demand. In addition, using different features without pre-filtering may have side effects on the accuracy. To determine the best combination of different features, ablation experiments were conducted in 4 steps.

1) Find the most sensitive feature to HAR and consider it the baseline. In this paper, there are 4 explainable features, and their variations are proposed. In this step, individual features and their variations were tested to determine the most sensitive feature to HAR. As shown in Fig. 7, feature C (relative coordinates) achieved the best performance and was consequently considered the baseline.

2) Add a new feature to the baseline, compare the accuracies of different new features, and select the best one as the new baseline.

3) Repeat step 2 until all features have been included. Then, compare all combinations and confirm the most cost-effective one. Tab. 3 records the detailed experiment results and shows that it is unnecessary to use all features in section 3.1. The combination of "$C + D$" achieved the equally best performance as the entire feature combination. Furthermore, "$C + D + \theta$" and " $C + D + D$" " performed worse than "$C + D$". Thus, an unsuitable feature may have side effects on the model, and more features are not necessarily better.

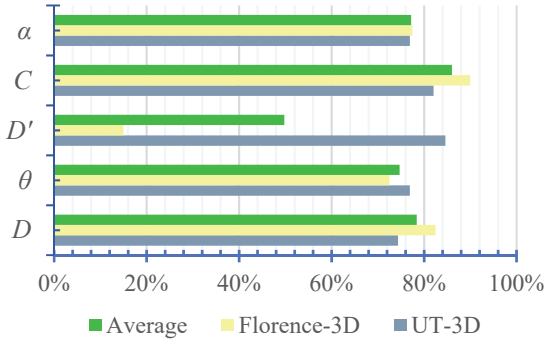


Figure 7 Accuracies of different features and their variations. Feature C (relative coordinates) achieved the best accuracy during the experiments, so it was selected as the baseline for the remaining experiments

4) Evaluate the global and local feature combinations. As a necessary complementary feature, the SMHI provides the motion tendency information.

Table 3 Comparison of the test accuracies of different CNNs

Baseline	Adding feature	Accuracy	
		UT-3d	Florence-3D
C	D	89.74%	90.00%
	θ	87.18%	87.50%
	D'	89.74%	60.00%
	α	89.74%	87.50%
$C + D$	θ	87.18%	90.00%
	D'	87.18%	90.00%
	α	89.74%	90.00%
$C + D + \alpha$	θ	87.18%	90.00%
	D'	89.74%	90.00%
$C + D + \alpha + D'$	θ	89.74%	90.00%

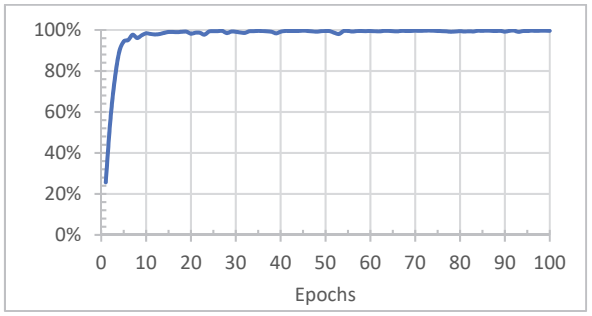
Tab. 4 shows that the SMHI plays an important role in all feature combinations. Even the single feature + SMHI method achieved satisfactory performance during the experiments. However, when a combination includes an increasing number of features, the effectiveness of the SMHI weakens. The reason is that the proposed features, which are derived from classical physics theories, contain important temporal information of the HAR, and the temporal information can partly reflect the motion tendency of the entire body. This finding also proves the importance and necessity of selecting explainable features.

To evaluate the proposed method, we adopted a more complex dataset that we collected. This dataset is named the HanYue-3D dataset, and it contains similar action groups and different temporal scales. Some actions that were manually arranged into different action groups are very difficult to discriminate. Fig. 8a and Fig. 8b show the accuracy curve and loss curve, respectively, during the training stage.

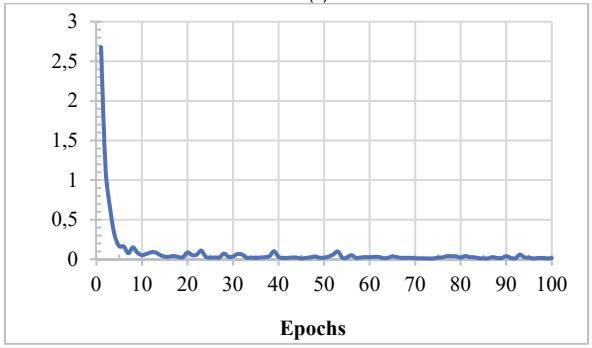
Table 4 Comparison of feature combinations with/without SMHIs.

Features	Accuracy	
	UT-3D	Florence-3D
C	82.05%	90.00%
C + SMHI	87.18%	92.50%
$C + D$	89.74%	90.00%
$C + D$ + SMHI	87.18%	92.50%
$C + D + \alpha$	89.74%	90.00%
$C + D + \alpha$ + SMHI	87.18%	92.50%
$C + D + \alpha + D'$	89.74%	90.00%
$C + D + \alpha + D'$ + *SMHI*	92.31%	92.50%
$C + D + \alpha + D'$	89.74%	90.00%
$C + D + \alpha + D' + \theta$ + SMHI	94.87%	92.50%

Figs. 8a and Fig. 8b)show that the curves remained stable after 20 epochs. This result proves that the model, explainable features and augmented training samples are effective and function well, and the 84.42% accuracy proves that the S-GMS can well address complex HAR problems.



(a)



(b)

Figure 8 Training curves of the S-GMS model tested on HanYue-3D. (a) Accuracy curve. (b) Loss curve

4.6 Comparison of the State-of-the-Art Methods

To continue evaluating the proposed method, we compared it with 4 typical CNN-based methods. The methods of literature [27, 28] are typical representatives of the joint coordinate-based methods, whereas the methods of literature [40, 41] are dataset collection methods. Selecting these 4 methods for comparison can effectively prove the superiority of our proposed method. The results of these methods are recorded in Tab. 5.

Table 5 Comparison with related CNN-based methods

Method	Dataset	Accuracy
Joint coordinates [27]	Florence-3D	67.50%
	UT-3D	82.05%
Arranged joint coordinates [28]	Florence-3D	70.00%
	UT-3D	79.49%
Dataset creators' method [40]	Florence-3D	82.20%
	UT-3D	
Dataset creators' method [41]	Florence-3D	
	UT-3D	90.92%
Our Method	Florence-3D	92.50%
	UT-3D	94.87%

Tab. 5 indicates at least 3 points.

1) Coordinate transformation is necessary and essential for feature extraction. Because the camera can be installed at any location, the direct usage of the original coordinates will inevitably result in a database-oriented problem. The proposed G-trans and L-trans strategies can effectively address this issue.

2) The use of the explainable feature is helpful for both action quantification and data augmentation. Compared with features that are extracted from a statistical viewpoint, the proposed features generated from a physical or kinematic viewpoint are more explainable. Thus, it is much easier to fully quantify the actions using limited features, and explainable features are much easier to find than a priori theories of physics and kinematics for data augmentation purposes.

3) Feature selection is more important than computing all features. The experiments effectively proved this point. When the number of features decreases, a deeper neural network is not better. A shallow network can efficiently achieve satisfactory performance.

The proposed method achieved the best performance in the experiments. Nonetheless, the skeleton-based method has inherent shortcomings. First, the best observation distance is 1.1 ~ 5 metres. Compared with the monocular camera, the observation distance is too short to be applied in outdoor environments. Second, the frequent monitoring deviation may cause all skeleton-based methods to have incorrect judgment. Finally, the estimated joint coordinates provided by the Kinect sensor always substantially deviate from their real values, which should be improved when some similar actions are differentiated.

5 CONCLUSIONS

This study presented a novel approach to HAR using explainable features and sparse motion history images. By carefully selecting a limited set of informative features and introducing the SMHI, we have demonstrated that efficient and accurate HAR can be achieved without relying on extensive feature sets. Our method outperforms several state-of-the-art approaches on three datasets while maintaining computational efficiency. The use of skeleton data ensures user privacy and makes our approach suitable for sensitive applications. Future work should focus on extending this method to more complex action sequences and exploring its potential in real-time applications. This research contributes to the ongoing effort to make HAR systems more interpretable, efficient, and practical for widespread deployment.

Acknowledgement

This research was supported by the Scientific and Technological Projects of the Nanchang Science and Technology Bureau under Grant GJJ212015, the National Natural Science Foundation of China under Grant 62366023, the Training Plan for Young Teachers in Key Disciplines of Jiangxi University of Chinese Medicine under Grant 2021jzzdxk021.

Authors contribution

Wei Yang: Methodology, Software development. Yitong Zhou: Data analysis, Validation. Shiwei Zhang and Lei Zhang: Funding acquisition. Jianying Xiong: Paper writing. Leiyue Yao: Supervision, Review & Editing.

Data availability and access

These lf-collected dataset, HanYue-3D, is available at: http://116.62.233.186:7777/HanYue-Action3D.zip

6 REFERENCES

[1] Wang, L., Huynh, D. Q., & Koniusz, P. (2019). A comparative review of recent Kinect-Based action recognition algorithms. *IEEE Transactions on Image Processing*, *29*, 15-28. https://doi.org/10.1117/12.2512989

[2] Sun, Z., Ke, Q., Rahmani, H., Bennamoun, M., Wang, G., & Liu, J. (2022). Human Action recognition from Various data Modalities: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1-20. https://doi.org/10.1109/TPAMI.2022.3183112

[3] Min, W., Yao, L., Lin, Z., & Liu, L. (2018). Suport vector machine approach to fall recognition based on simplified expression of human skeleton action and fast detection of start key frame using torso angle. *IET Computer Vision*, *12*(8), 1133-1140. https://doi.org/10.1049/iet-cvi.2018.5324

[4] Bobick, A. & Davis, J. (2001). The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(3), 257-267. https://doi.org/10.1109/34.910878

[5] Zhang, S., Chen, E., Qi, C., & Liang, C. (2016). Action Recognition Based on Sub-action Motion History Image and Static History Image. *MATEC Web of Conferences*, *56*, 02006. https://doi.org/10.1051/matecconf/20165602006

[6] Weinland, D., Ronfard, R., & Boyer, E. (2006). Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding*, *104*(2-3), 249-257. https://doi.org/10.1016/j.cviu.2006.07.013

[7] Laptev, I. (2005). On Space-Time Interest Points. *International Journal of Computer Vision*, *64*(2-3), 107-123. https://doi.org/10.1007/s11263-005-1838-7

[8] Dalal, N. & Triggs, B. (2005). Histograms of oriented gradients for human detection. *2005 IEEE Computer Society*

Conference on Computer Vision and Pattern Recognition (*CVPR'05*), *1*, 886-893. https://doi.org/10.1109/CVPR.2005.177

[9] Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. *IEEE Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/CVPR.2008.4587756

[10] Wang, Y. & Sun, J. (2022). Video Human Action Recognition Algorithm Based on Double Branch 3D-CNN. *2022 15th International Congress on Image and Signal Processing*, *BioMedical Engineering and Informatics* (*CISP-BMEI*). https://doi.org/10.1109/CISP-BMEI56279.2022.9979858

[11] Wang, R., Luo, H., Wang, Q., Li, Z., Zhao, F., & Huang, J. (2020). A Spatial–Temporal Positioning Algorithm Using Residual Network and LSTM. *IEEE Transactions on Instrumentation and Measurement, 69*(11), 9251-9261. https://doi.org/10.1109/TIM.2020.2998645

[12] Li, H., Shrestha, A., Heidari, H., Kernec, J. L., & Fioranelli, F. (2019). Bi-LSTM Network for Multimodal Continuous Human Activity Recognition and Fall Detection. *IEEE Sensors Journal*, *20*(3), 1191-1201. https://doi.org/10.1109/JSEN.2019.2946095

[13] Elaoud, A., Ghazouani, H., & Barhoumi, W. (2024). XYZ-channel encoding and augmentation of human joint skeleton coordinates for end-to-end action recognition. *Signal Image and Video Processing*. https:// doi.org/10.1007/s11760-024-03434-4

[14] Surendran, R., J, A., & Hemanth, J. D. (2023). Recognition of human action for scene understanding using world cup optimization and transfer learning approach. *Peer J Computer Science*, *9*, e1396. https://doi.org/10.7717/peerj-cs.1396

[15] Bilal, M., Maqsood, M., Yasmin, S., Hasan, N. U., & Rho, S. (2021). A transfer learning-based efficient spatiotemporal human action recognition framework for long and overlapping action classes. *The Journal of Supercomputing*, *78*(2), 2873-2908. https://doi.org/10.1007/s11227-021-03957-4

[16] Lin, K., Zhou, J., & Zheng, W. (2024). Human-Centric Transformer for Domain Adaptive Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1-18.

[17] Yang, N. Y., Saleemi, I., & Shah, M. (2012). Discovering Motion Primitives for Unsupervised Grouping and One-Shot Learning of Human Actions, Gestures, and Expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(7), 1635-1648. https://doi.org/10.1109/TPAMI.2012.253

[18] Li, Z., Gong, X., Song, R., Duan, P., Liu, J., & Zhang, W. (2022). SMAM: self and mutual adaptive matching for skeleton-based few-shot action recognition. *IEEE Transactions on Image Processing*, *32*, 392-402. https://doi.org/10.1109/TIP.2021.3130533

[19] Ji, Z., Liu, X., Pang, Y., Ouyang, W., & Li, X. (2020). Few-Shot Human-Object Interaction Recognition With Semantic-Guided Attentive Prototypes Network. *IEEE Transactions on Image Processing*, *30*, 1648-1661. https://doi.org/10.1609/aaai.v34i07.6764

[20] Zhao, Y., Shi, P., & You, J. (2019). Fine-grained Human Action Recognition Based on Zero-Shot Learning. *IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)*, *26*, 294-297. https://doi.org/10.1109/ICSESS47205.2019.9040818

[21] Sato, F., Hachiuma, R., & Sekii, T. (2023). Prompt-Guided Zero-Shot Anomaly Action Recognition using Pretrained Deep Skeleton Features. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). https://doi.org/10.1109/CVPR52729.2023.00626

[22] Cheung, T. & Yeung, D. (2023). A Survey of Automated Data Augmentation for Image Classification: Learning to Compose, Mix, and Generate. *IEEE Transactions on Neural Networks and Learning Systems*, 1-21. https://doi.org/10.1109/TNNLS.2023.3282258

[23] Khan, S., Hassan, A., Hussain, F., Perwaiz, A., Riaz, F., Alsabaan, M., & Abdul, W. (2023). Enhanced Spatial Stream of Two-Stream Network Using Optical Flow for Human Action Recognition. *Applied Sciences*, *13*(14), 8003. https://doi.org/10.3390/app13148003

[24] Giveki, D. (2024). Human action recognition using an optical flow-gated recurrent neural network. *International Journal of Multimedia Information Retrieval*, *13*(3). https://doi.org/10.1007/s13735-024-00338-4

[25] Berlin, S. J. & John, M. (2020). Spiking neural network based on joint entropy of optical flow features for human action recognition. *The Visual Computer*, *38*(1), 223-237. https://doi.org/10.1007/s00371-020-02012-2

[26] Chen, H., Wang, G., Xue, J., & He, L. (2016). A novel hierarchical framework for human action recognition. *Pattern Recognition, 55*, 148-159. https://doi.org/10.1016/j.patcog.2016.01.020

[27] Du, Y., Fu, Y., & Wang, L. (2015). Skeleton based action recognition with convolutional neural network. *2015 3rd IAPR Asian Conference on Pattern Recognition* (*ACPR*), 579-583. https://doi.org/10.1109/ACPR.2015.7486569

[28] Yang, Z., Li, Y., Yang, J., & Luo, J. (2018). Action Recognition with Spatio-Temporal Visual Attention on Skeleton Image Sequences. *IEEE Transactions on Circuits and Systems for Video Technology*, *29*(8), 2405-2415. https://doi.org/10.1109/TCSVT.2018.2864148

[29] Yao, L., Yang, W., & Huang, W. (2020). A data augmentation method for human action recognition using dense joint motion images. *Applied Soft Computing*, *97*, 106713. https://doi.org/10.1016/j.asoc.2020.106713

[30] Yao, L., Yang, W., Huang, W., Jiang, N., & Zhou, B. (2021). Multi-scale feature learning and temporal probing strategy for one-stage temporal action localization. *International Journal of Intelligent Systems*, *37*(7), 4092-4112. https://doi.org/10.1002/int.22713

[31] Liao, L., Yang, Y., & Fu, L. (2019). Joint-oriented Features for Skeleton-based Action Recognition. *IEEE International Conference on Systems, Man and Cybernetics (SMC)*, *15*, 1154-1159. https://doi.org/10.1109/SMC.2019.8914565

[32] Dhiman, C., Saxena, M., & Vishwakarma, D. K. (2019). Skeleton-Based View Invariant Deep Features for Human Activity Recognition. *2019 IEEE Fifth International Conference on Multimedia Big Data* (*BigMM*), 225-230. https://doi.org/10.1109/BigMM.2019.00-21

[33] Benhamida, L. & Larabi, S. (2024). Human action recognition using ST-GCNs for blind accessible theatre performances. *SignalImage and Video Processing*. https://doi.org/10.1007/s11760-024-03510-9

[34] Enkhbat, A., Shih, T. K., & Cheewaprakobkit, P. (2024). Human Action Recognition and Note Recognition: A Deep Learning Approach Using STA-GCN. *Sensors*, *24*(8), 2519. https://doi.org/10.3390/s24082519

[35] Chi, H., Ha, M. H., Chi, S., Lee, S. W., Huang, Q., & Ramani, K. (2022). InfoGCN: Representation Learning for Human Skeleton-based Action Recognition. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*). https://doi.org/10.1109/CVPR52688.2022.01955

[36] Planinc, R. & Kampel, M. (2012). Introducing the use of depth data for fall detection. *Personal and Ubiquitous Computing*, *17*(6), 1063-1072. https://doi.org/10.1007/s00779-012-0552-z

[37] Yao, L., Min, W., & Lu, K. (2017). A new approach to fall detection based on the human torso motion model. *Applied Sciences*, *7*(10), 993-1009. https://doi.org/10.3390/app7100993

[38] Wang, Y. & Sun, J. (2022). Video Human Action Recognition Algorithm Based on Double Branch 3D-CNN. *2022 15th International Congress on Image and Signal Processing*, *BioMedical Engineering and Informatics (CISP-BMEI)*. https://doi.org/10.1109/cisp-bmei56279.2022.9979858

[39] Li, J., Liu, X., Zhang, M., & Wang, D. (2019). Spatio-temporal deformable 3D ConvNets with attention for action recognition. *Pattern Recognition*, *98*, 107037. https://doi.org/10.1016/j.patcog.2019.107037

[40] Seidenari, L., Varano, V., Berretti, S., Del Bimbo, A., & Pala, P. (2013). Recognizing Actions from Depth Cameras as Weakly Aligned Multi-Part Bag-of-Poses.*2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 479-485. https://doi.org/10.1109/CVPRW.2013.77

[41] Xia, L., Chen, C., & Aggarwal, J. K. (2012). View invariant human action recognition using histograms of 3D joints. *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshop*, 20-27. https://doi.org/10.1109/CVPRW.2012.6239233

Contact information:

Wei YANG, Associate professor
Jiangxi University of Technology,
The Center of Collaboration and Innovation,
Nanchang, China 330098
E-mail: Wei.yang@163.com

Yitong ZHOU
Jiangxi University of Chinese Medicine,
The College of Computer Science,
Nanchang, China 330004
E-mail: 13576255881@163.com

Jianying XIONG, Associate professor
Jiangxi University of Chinese Medicine,
The College of Computer Science,
Nanchang, China 330004
E-mail: special8212@sohu.com

Shiwei ZHANG
The Hanlin Hangyu (Tianjin) Industrial Co.,
Ltd. Tianjin, China 301899
E-mail: shiwei.zhang@bjhanlin.com

Lei ZHANG
The Hanlin Hangyu (Tianjin) Industrial Co.,
Ltd. Tianjin, China 301899
E-mail: sanshilei@126.com

Leiyue YAO, Professor
(Corresponding author)
Jiangxi University of Chinese Medicine,
The College of Computer Science,
Nanchang, China 330004
E-mail: Leiyue_yao@163.com