HUMAN ACTION RECOGNITION METHOD BASED ON HIERARCHICAL FRAMEWORK VIA KINECT SKELETON DATA

BENYUE SU[1,2]**, HUANG WU**[1,2]**, MIN SHENG**[2,3]

[1]School of Computer and Information, Anqing Normal University, Anqing 246133, China
[2]The Key Laboratory of Intelligent Perception and Computing of Anhui Province, Anqing 246133, China
[3]School of Mathematics and Computational Science, Anqing Normal University, Anqing 246133, China
E-MAIL: bysu@aqnu.edu.cn, vic.woo@qq.com, msheng0125@aliyun.com

Abstract:

Human action recognition is a hot issue in the field of machine vision. It play a pivotal role in human-centered computing. There are challenges mainly from the complexity of human actions and and high-noise data. Here needs to solve problems such as high intra-class variance with low inter-class variance, variable movement speed, and high computational costs. Based on the above points, we use a thought of hierarchy to design a multi-level hierarchical recognition model. Owing to the expression of the systemic actions and the local actions are different, so different features are used at different levels. The method of hierarchical classification use proper classification algorithm in different levels, to subdivide category layer by layer, until it cannot be subdivided. In this paper, we use a two-level hierarchical framework based on the MSRAction3D dataset using skeleton data which captured via Kinect sensor. At the first level, we use Support Vector Machine to classify all categories into seven categories. At the second level, we use the Hidden Markov Model to reclassify seven categories. Experimental results show that our method is superior to other state-of-the-art methods, achieving 91.41% average recognition rate. The idea of stratification is applied to human action recognition can embodies the inherent level relationship of human movement.

Keywords:

Human action recognition; Multi-level hierarchical recognition model; Skeleton data; Kinect; Support Vector Machine; Hidden Markov Model

1. Introduction

Human action recognition (hereafter referred to as HAR) is a hot research topic in field of computer vision for decades [1]. It is widely used in human-computer interaction, virtual reality, video surveillance, medical rehabilitation and other fields. In recent years, domestic and foreign research on HAR has made important progress. There are two main directions for HAR at the current study. Optical human action recognition and Wearable human action recognition. Since wireless sensing network technology has improved quickly, Wearable human action recognition, Wearable human motion recognition has developed rapidly in recent years [2]. However, due to its high cost, the human body is easy to be fettered, wearing inconvenience and other shortcomings, HAR based on wearable devices still can't shake the dominant position of optical devices in the direction of HAR. Because of its low cost, convenient operation, no influence on human body and good user experience, HAR based on optical devices has always been the main research content in the field of HAR. In particular, a new generation of optical sensor called Kinect released by Microsoft (see in Figure 1). It brings new vitality to the research of optical human action recognition.

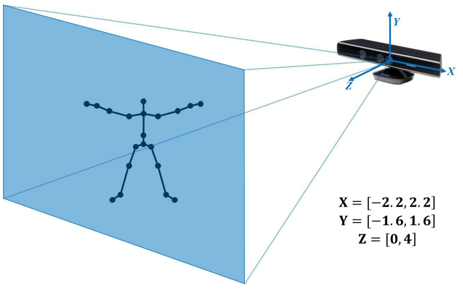


FIGURE 1. Person in Kinect

Kinect has 3 cameras altogether, the middle one is the RGB camera and 2 depth sensors on both sides, the infrared transmitter is on the left, and the infrared receiver is on the right [3]. Kinect for Windows SDK is Microsoft's official tool for

978-1-5386-0408-3/17/$31.00 ©2017 IEEE

many developers to research, test and experiment. The greatest feature of Kinect is Skeleton tracking. As shown in Figure 1, person moving in the Kinect field of view, Skeleton tracking technology can be directly access to human skeleton movement information through its own SDK [4]. In this way, we obtain skeleton data that can represent human actions.

In this paper, we use skeleton data via kinect for HAR. The highlights of this paper are the follows:

- Multi-granularity attribute based on human action, we use hierarchical theory to design a multi-level hierarchical recognition model, for achieving the classification of human action by layer.
- We use cascade classifier to select proper classification algorithm in different levels with different features, for subdividing category layer by layer, until it cannot be subdivided.

The paper is organized as follows: The Section 1 of this paper introduces the background of HAR based on skeleton data via kinect. In Section 2, it describes the background and significance of the hierarchical framework. In Section 3, it explains the feature descriptor, extracting different features are used at different levels in the hierarchical recognition model. In Section 4, it introduces the construction and usage of classifier. Next, we present our experiment method and display experiment results in Section 5. Finally, There is a brief summary is given in Section 6.

2. Hierarchical Framework

The greatest challenge for HAR is the effective recognition of human actions. The source of the challenges are mainly from two points the follows: the complexity of human actions and difficulties are inherent in data [5].

Human action is a complex process. There are many factors that affect human actions such as the environment, culture, personal differences, and emotions. The large diversity of human body size, appearance, and shape. Different people will perform the same action differently, and even the same person will perform it differently at different times. Kinect is essentially an optical sensor, it also has the inherent problems of optical sensors such as action occlusion, resulting in missing data. The data are captured in different environmental settings, with different camera settings, and fixed or dynamic camera positions. Moreover, owing to kinect v1.0 uses structured light to obtain depth data, the data are captured by kinect behaves very high noise. Therefore, action recognition using kinect need to overcome some issues such as high intra-class variance with low inter-class variance, variable movement speed, and high computational costs. According to general knowledge, we know that human exercise is divided into systemic exercise and local exercise. The movement of the body is always accompanied by changes from a wide range of movement to a small range of motion. That is to say, body movement presents a process of gradual change from coarse-grained to fine-grained. This is a process of stepwise precise classification layer by layer. Based on the above points, we use a thought of hierarchy to design a multi-level hierarchical recognition model. Owing to the expression of the systemic actions and the local actions are different, so different features are used at different levels in the hierarchical recognition model. The method of hierarchical classification use proper classification algorithm in different levels, to subdivide category layer by layer, until it cannot be subdivided.

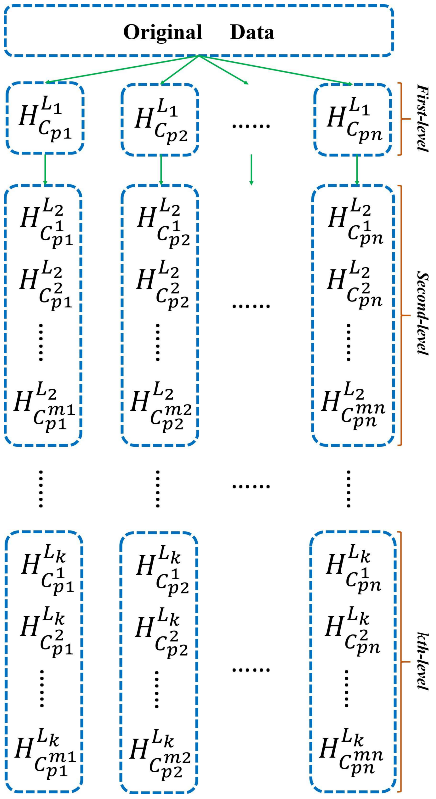


FIGURE 2. Hierarchical framework for human action recognition

As shown in Figure 2, the original class contains many types of action data, according to the crudeness or fineness of the action granularity, stratifying the action data. $H^{L_k}_{C^{mn}_{pn}}$ is explained below :

- $H_{C_{pn}^{mn}}^{L_k}$ itself represents a category.
- Superscript L_k represents the model is the k-level.
- Subscript C_{pn}^{mn} represents the mnth subcategories of pnth category .
- $s.t. \quad k, pn, mn \in N^*$.

In this article, we use a two-level hierarchy based on the public dataset that we choose. The first level of recognition is divided into several classes based on which parts of the human body to participate in action. The second level of recognition is the subdivision of the previous action category based on the same parts of the human body to participate in action.

3. Feature Extraction

There are two key steps in HAR, the first step is the extraction of action features. Kinect has a special function that can directly access the location information of human skeleton. The skeleton tracking technique establishes the coordinates of the joints of the human body by processing the depth data. It also can determine the parts of the body, such as the hand, the head, and the body, and the location where they are located. With the kinect SDK, we can obtain the coordinates of the 20 human skeleton points as shown in Figure 3.

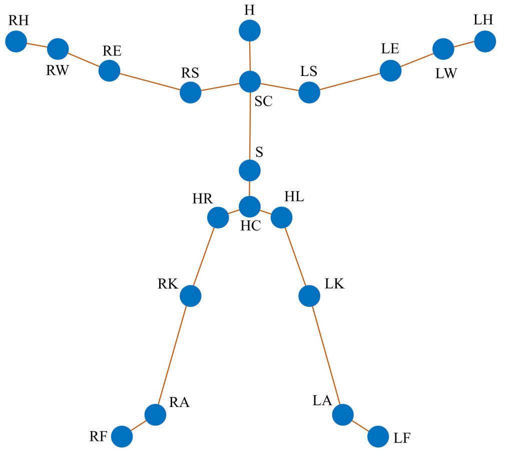
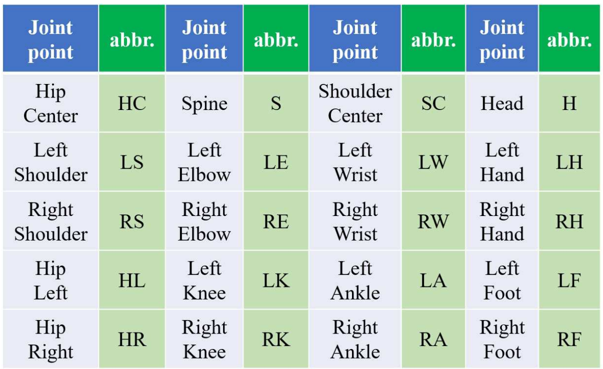


FIGURE 3. 20 skeleton points obtained by Kinect v1.0

In order to facilitate understanding, we made the table as shown in Figure 4.



Joint point	abbr.	Joint point	abbr.	Joint point	abbr.	Joint point	abbr.
Hip Center	HC	Spine	S	Shoulder Center	SC	Head	H
Left Shoulder	LS	Left Elbow	LE	Left Wrist	LW	Left Hand	LH
Right Shoulder	RS	Right Elbow	RE	Right Wrist	RW	Right Hand	RH
Hip Left	HL	Left Knee	LK	Left Ankle	LA	Left Foot	LF
Hip Right	HR	Right Knee	RK	Right Ankle	RA	Right Foot	RF

FIGURE 4. The name of 20 skeleton joint points

Suppose that there are I joint points (the human skeleton points are extracted by kinect v1.0, so $I = 20$ in this paper) represent a human body, and the action of human is performed over T frames. Therefore, let $x_i^{(t)}$, $y_i^{(t)}$ and $z_i^{(t)}$ be the X, Y, and Z coordinates of the i^{th} joint point at frame t. Therefore, we can use this formula $P_i^{(t)} = (x_i^{(t)}, y_i^{(t)}, z_i^{(t)})$ $(i = 1, 2, 3, \ldots, I; t = 1, 2, 3, \ldots, T)$ to represent the spatio-temporal position of each joint point. For example, we express the coordinates of the right hand at frame t, as follows :

$$P_{RH}^{(t)} = (x_{RH}^{(t)}, y_{RH}^{(t)}, z_{RH}^{(t)}) \tag{1}$$

Note that $P_{RH}^{(t)}$ in formula (1), P means joint point, RH means abbreviation of joint point name replace i for convenience. Similarly, we can use this method to represent other joint points. In this way, we divide the human body into five parts based on human anatomy.

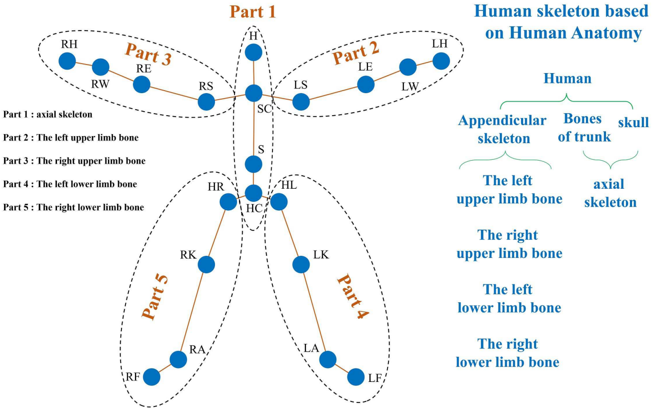

Part 1
Part 2
Part 3
Part 4
Part 5
RH
RE
RW
RS
H
SC
S
LS
LE
LW
LH
HR
HC
HL
RK
LK
RA
LA
RF
LF
Part 1 : axial skeleton
Part 2 : The left upper limb bone
Part 3 : The right upper limb bone
Part 4 : The left lower limb bone
Part 5 : The right lower limb bone
Human skeleton based on Human Anatomy
Human
Appendicular skeleton
Bones of trunk
skull
The left upper limb bone
The right upper limb bone
The left lower limb bone
The right lower limb bone
axial skeleton

FIGURE 5. Skeleton classification based on human anatomy

As shown in Figure 5, the first part represents axial skeleton, the second part represents left-upper-extremity skeleton, the third part represents right-upper-extremity skeleton, the fourth part represents left-lower-extremity skeleton and the fifth part represents right-lower-extremity skeleton. Specifically, as shown below :

$$\begin{aligned} Part\,1 &: \{P_{HC}^{(t)}, P_{S}^{(t)}, P_{SC}^{(t)}, P_{H}^{(t)}\} \\ Part\,2 &: \{P_{LS}^{(t)}, P_{LE}^{(t)}, P_{LW}^{(t)}, P_{LH}^{(t)}\} \\ Part\,3 &: \{P_{RS}^{(t)}, P_{RE}^{(t)}, P_{RW}^{(t)}, P_{RH}^{(t)}\} \\ Part\,4 &: \{P_{HL}^{(t)}, P_{LK}^{(t)}, P_{LA}^{(t)}, P_{LF}^{(t)}\} \\ Part\,5 &: \{P_{HR}^{(t)}, P_{RK}^{(t)}, P_{RA}^{(t)}, P_{RF}^{(t)}\} \end{aligned} \tag{2}$$

Certainly, the data should be processed to eliminate the effects of individual differences. we choose the distance between the shoulders as a normalized scale in (3), because the distance between the shoulders in the course of movement remained essentially unchanged.

$$|P_{LS}^{(t)} - P_{RS}^{(t)}| \tag{3}$$

Without considering the case of the fine actions such as the finger participation. Human actions can be effectively recognized using a two-level hierarchical framework.

At the first level, our goal is to distinguish which parts of the human body are moving over time frames. We use the barycenter of the first three joint points of each part to represent the part. For example, suppose we use $C_1^{(t)}$ to represent the barycenter of the first part of the body at frame t. Because the human body has been divided into 5 parts, it is necessary to construct the local coordinate system in coordinate calculation. Then its mathematical expression is as follows:

$$\begin{aligned} C_1^{(t)} &= \frac{P_{HC}^{(t)} + P_{S}^{(t)} + P_{SC}^{(t)}}{3} - P_{HC}^{(t)} \\ \Longrightarrow\ & C_1^{(t)} = (x_{C_1}^{(t)}, y_{C_1}^{(t)}, z_{C_1}^{(t)}) \end{aligned} \tag{4}$$

In the same way, we can get the $C_2^{(t)}$, $C_3^{(t)}$, $C_4^{(t)}$, $C_5^{(t)}$. and then, the distance between the frame and the initial frame on the X axes, Y axes and Z axes is calculated from the second frame. Furthermore, calculating respective mean and variance.

Taking the X axis as an example, the formula is as follows. $\tilde{x}_{C_1}$ represents the distance change.

$$\begin{aligned} &\tilde{x}_{C_1}^{(t-1)} = ||x_{C_1}^{(t)} - x_{C_1}^{(1)}||^2 \\ &Mean\ :\ E(\tilde{x}_{C_1}) = \frac{\sum_{t=2}^{T} \tilde{x}_{C_1}^{(t-1)}}{T-1} \\ &Variance\ :\ D(\tilde{x}_{C_1}) = \frac{\sum_{t=2}^{T} ||\tilde{x}_{C_1}^{(t-1)} - E(\tilde{x}_{C_1})||^2}{T-1} \\ &s.t. \quad t = 2, 3, \cdots, T \end{aligned} \tag{5}$$

Note that $x_{C_1}^{(1)}$ in formula (5), represents the calibration frame of the action. In the same way, we can get the rest of statistical features such as $E(\tilde{y}_{C_1}), E(\tilde{z}_{C_1}), D(\tilde{y}_{C_1}), D(\tilde{z}_{C_1})$. Similarly, the statistical features of the remaining four parts are also available. we extract the feature vector of the first level in this way. Finally, we can get the first-level feature vector $FV1$:

$$\begin{aligned} &FV1 = (FV1_{C_1}, FV1_{C_2}, FV1_{C_3}, FV1_{C_4}, FV1_{C_5}) \\ &FV1_{C_i} = (E(\tilde{x}_{C_i}), E(\tilde{y}_{C_i}), E(\tilde{z}_{C_i}), D(\tilde{x}_{C_i}), D(\tilde{y}_{C_i}), D(\tilde{z}_{C_i})) \\ &s.t. \quad i = 1, 2, 3, 4, 5 \end{aligned} \tag{6}$$

At the second level, we need to classify the similar actions which have same parts involved in. Action is a sequence of temporal and spatial relationships, so we need to show the time attribute of each action. In the first level, we use the first three joint points of each part to extract features, and in the second level, we will use the fourth joint point. It is the end-effector of the human body action, its movement trajectory can best reflect the features of the human action. Next, let's take the third part (RS, RE, RW, RH) as an example. In order to facilitate mathematical representation, we use E to represent the end-effector. Therefore, the third part end-effector (RH) at current frame t can be expressed as $E_3^{(t)}$, the rest and so on. So, the calculating formula of local position of the third part end-effector is as follows:

$$\begin{aligned} &E_3^{(t)} = P_{RH}^{(t)} - P_{RS}^{(t)} \\ \Longrightarrow\ & E_3^{(t)} = (x_{E_3}^{(t)}, y_{E_3}^{(t)}, z_{E_3}^{(t)}) \end{aligned} \tag{7}$$

In human kinematics, the position of each moment can reflect the movement of the object in space, that is, the trajectory of movement. And the location changes before and after the time can reflect the movement of objects in time, that is, the speed of movement. Calculating the relative offset distance on the X axes, Y axes and Z axes between the next frame $t+1$ and the current frame t :

$$v(E_3^{(t)}) = (v(x_{E_3}^{(t)}), v(y_{E_3}^{(t)}), v(z_{E_3}^{(t)}))$$

$$\begin{cases} v(x_{E_3}^{(t)}) = & x_{E_3}^{(t+1)} - x_{E_3}^{(t)} \\ v(y_{E_3}^{(t)}) = & y_{E_3}^{(t+1)} - y_{E_3}^{(t)} \quad t = 1, 2, \cdots, T-1 \\ v(z_{E_3}^{(t)}) = & z_{E_3}^{(t+1)} - z_{E_3}^{(t)} \end{cases} \tag{8}$$

In the above formula, $E_3^{(t)}$ is Coordinate representation of right hand in local coordinate system the current frame. $v(E_3^{(t)})$ actually represents the speed of the current frame relative to the previous frame. Therefore, we believe that E_3, $v(E_3)$ can reflect the the space-time attributes of human action. It can be used as features of human action recognition. which as shown in Figure 6.

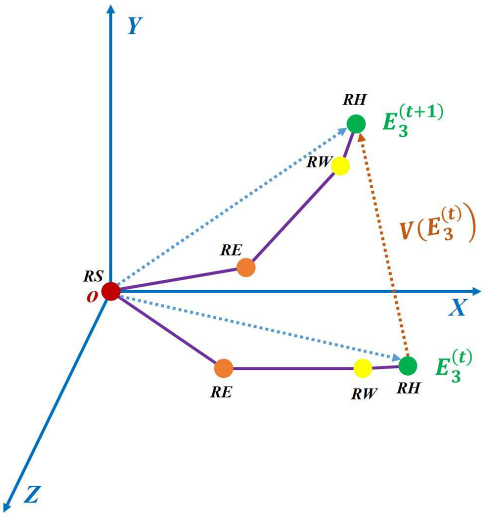


FIGURE 6. The second-level features of Right Arm

Similarly, Other parts of the end effector's features, and so on. Finally, we can get the second-level feature vector $FV2$, the formula is as follows :

$$FV2 = (FV2_{E_1} \oplus FV2_{E_2} \oplus FV2_{E_3} \oplus FV2_{E_4} \oplus FV2_{E_5})$$

$$FV2_{E_j} = (x_{E_j}, y_{E_j}, z_{E_j}, v(x_{E_j}), v(y_{E_j}), v(z_{E_j}))$$

$$s.t. \quad j = 1, 2, 3, 4, 5 \tag{9}$$

Finally, we need to standardize the feature vector before entering the classifier. So that the data dimensionless to eliminate the differences caused by different units. After that, the next step, we will select the appropriate classifier to identify the human actions, according to the nature of each level features and the characteristics of the human body. Classifier Construction is below.

4. Classifier Construction

Human action recognition is difficult to recognize because of the complexity of its actions. In order to overcome this difficulty, we design a two-level hierarchical framework recognition model based on the knowledge of the human anatomy and the human kinematics. Generally, we give the algorithm flow-process diagram of experiment as shown in Figure 7. The solid line represents the input path of the training set, and the dashed line represents the input path of the test set.

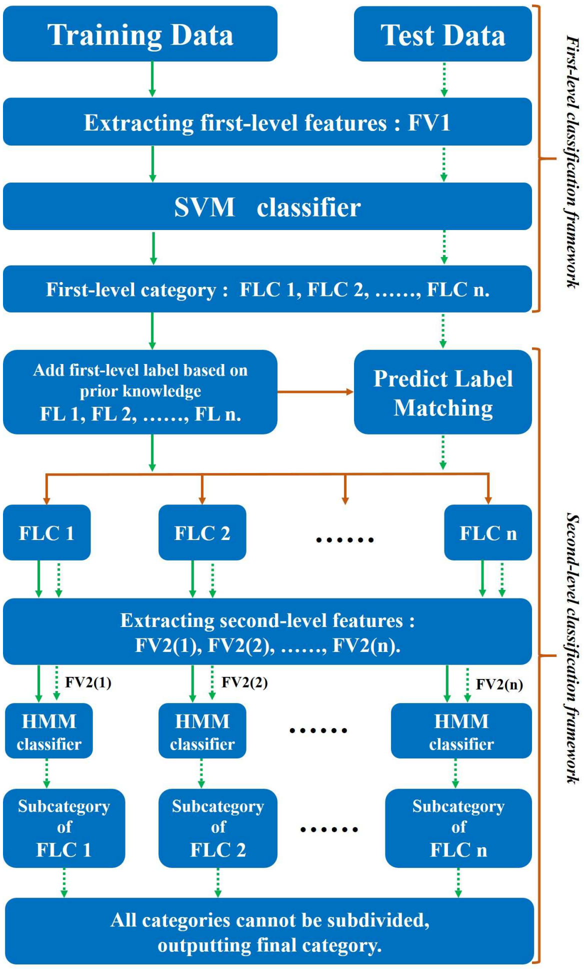

Training Data
Test Data
Extracting first-level features : FV1
SVM classifier
First-level category : FLC 1, FLC 2, ……, FLC n.
First-level classification framework
Add first-level label based on prior knowledge FL 1, FL 2, ……, FL n.
Predict Label Matching
FLC 1
FLC 2
……
FLC n
Extracting second-level features : FV2(1), FV2(2), ……, FV2(n).
FV2(1)
FV2(2)
FV2(n)
HMM classifier
HMM classifier
……
HMM classifier
Subcategory of FLC 1
Subcategory of FLC 2
……
Subcategory of FLC n
All categories cannot be subdivided, outputting final category.
Second-level classification framework

FIGURE 7. Algorithm flow-process of two-level recognition model

At the first level, we divide the body into five parts. According to the five parts, whether to participate in the movement, make the coarse classification. For example, if only third parts of the human body(RS, RE, RW, RH) had action, while the rest of the parts remained stationary, then it is classified as a class, for the convenience of introduction, which is called the right arm action. Similarly, if both the second part(LS, LE, LW, LH) and the third part(RS, RE, RW, RH) had action at the same time, then it is called the double arm action, and so on. This classification is in line with the human anatomy and the normal human cognition. In this paper, our experiments are based on MSRAction3D dataset. At the first level, we need to divide the 20 actions into several categories. There are many different actions that use the same parts in the dataset. As shown in Figure 8, We can divide the current data sets into seven categories, and then classify them into smaller classes.

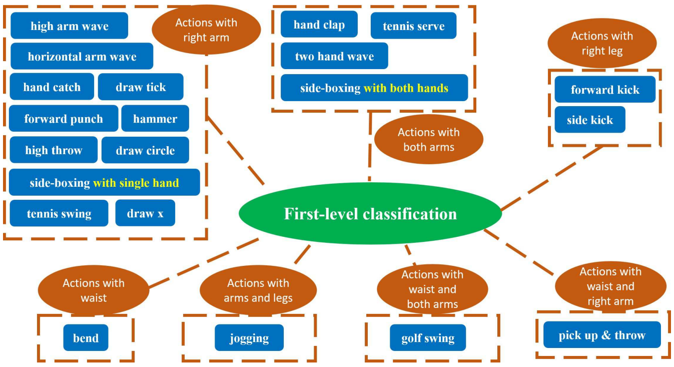


FIGURE 8. First-level classification based on MSRAction3D

At the second level, we need to classify the actions of the same parts. This is a fine classification that is different from the previous rough classification. At this level, the classification of the action should focus on the local action features, the features should be more detailed description of the action itself. Therefore, for different parts should choose the appropriate features and classifiers to recognition of human actions. The detailed algorithm process is shown below. At the second level, according to prediction labels based on first layer classification, we designed a judgment mechanism. If there are subcategories in the category corresponding to the predict label, the category is continually classified and recognized until the final predict label is the original category. For different feature attribute, we choose different classifier. There is two kinds of feature type, feature without temporal relationship such as mean and variance and feature with temporal relationship such as time series. The former we named static data, the latter we call dynamic data. We use SVM for static data, and we use HMM for dynamic data.

At the first-level of classification, we use the features called $FV1$ (See formula 6) with SVM. Because of the first level recognition is rough classification, the action categories of the first level are very different, which has large intra-class distance and small inter-class variance can achieve high recognition performance. Therefore, mean and variance of human action over time can be used as features for SVM effective classification recognition. And mean and variance without temporal relationship, it belongs to static data. SVM for static data also has a good classification effect.

At the second-level of classification, we get several subcategories through the first-level classifier. The subcategories are have small intra-class distance and large inter-class variance, therefore, it need for more detailed features for the fine classification. For different subcategories, we choose proper different features. Because the second level action is related to the local action of the human body. So we need to choose features according to the different parts which involved in action. For example, we divide the all actions in MSRAction3D dataset into seven major categories at First-level classification. Actions with right arm have 11 subcategories, actions with both arms have 4 subcategories and actions with right leg have 2 subcategories. rest of actions are only one, no need to break down. For that three major categories we use they are features which part has involved in the action to to classify. As shown in Figure 9, we use $FV2_{(1)}$ for actions with right arm, $FV2_{(2)}$ for actions with both arms, $FV2_{(3)}$ for actions with right leg, and so on.

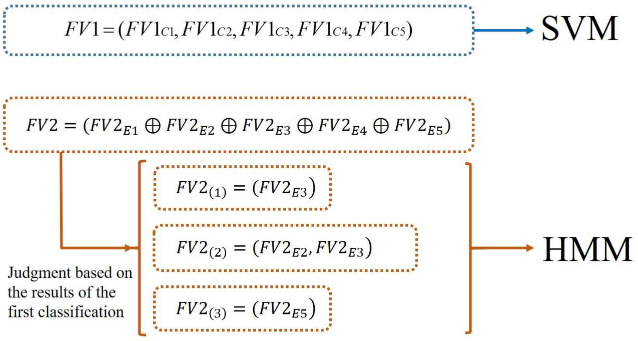

$FV1 = (FV1_{C1}, FV1_{C2}, FV1_{C3}, FV1_{C4}, FV1_{C5})$
SVM
$FV2 = (FV2_{E1} \oplus FV2_{E2} \oplus FV2_{E3} \oplus FV2_{E4} \oplus FV2_{E5})$
$FV2_{(1)} = (FV2_{E3})$
$FV2_{(2)} = (FV2_{E2}, FV2_{E3})$
$FV2_{(3)} = (FV2_{E5})$
Judgment based on the results of the first classification
HMM

FIGURE 9. Features and classifiers based on MSRAction3D

We find that there subcategories are actions with parts of human body are high similarity, The description of the action process is highly dependent. Because HMM can express a state to another state transition process, and the human action details can be expressed as a state to another state changes. HMM has effective recognition on dynamic time series data, $FV2$ (See formula 9) just is a feature over time. Therefore, the second level features are very suitable for classification using HMM.

5 Experiment

In this paper, the hardware setup : Intel Core i5 4210M CPU @ 2.6 GHz, 8 GB RAM. And the software setup : Windows 10 64-bit operating system and MATLAB R2015b. The public dataset: MSRAction3D Dataset.

MSRAction3D consists of 20 action types of 10 subjects, each subject performs each action 2 or 3 times. There are 567 action sequence files in total. Note that there is an error in the paper on the number of samples being used for the experiment. The number 402 in the paper is not correct. The correct number is 557. Out of the original 567 sequences in MSR Action3D Dataset, 10 sequences are not used in this paper's experiment because the skeletons are either missing or too erroneous [6].

MSRAction3D contains twenty actions: ***high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pick up & throw***. The action types in the dataset are shown in Figure 10.

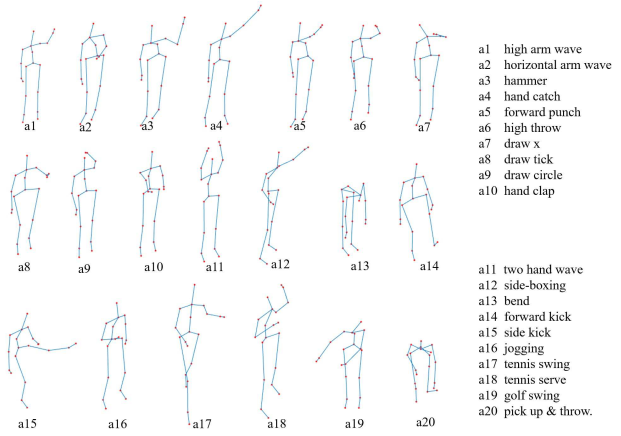


FIGURE 10. Display action type in MSRAction3D

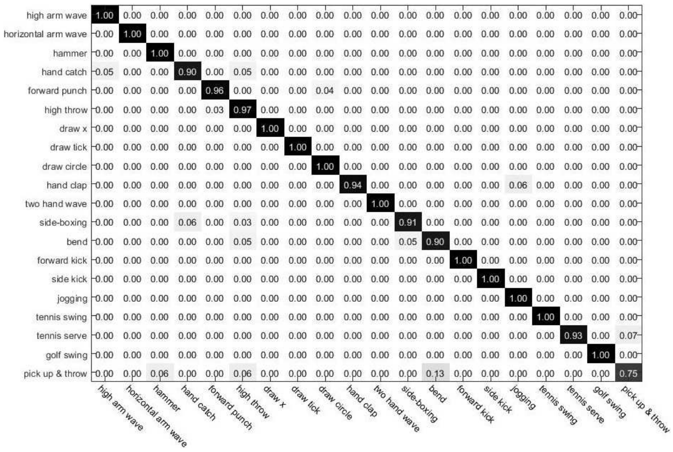
In the most traditional approach based on MSRAction3D, as with all the action sequences of the dataset are split into three subsets, each with eight actions. Experiments in this paper, we use entire 20 actions with 547 sequences are applied (remove 20 samples of serious data loss).

In the experiment, K-fold cross validation (K-CV) method was used to select test samples and training samples. During the experiment, the dataset is divided into 10 groups (the samples data in MSRAction3D dataset are from 10 different subjects. Therefore, the K value is 10, that is to say, 10-fold cross validation, and the recognition rate independent of the individual sample can be obtained by this method). Each group was selected as the test set and the remaining rest groups were used as training set. In accordance with the above method to do ten times, and the average is taken as a result finally. A total of 100 times in our experiment.

TABLE 1. Recognition rates experiment on MSRAction3D.

Methods (Recognition rate unit:%)	*Best*	*Worst*	*Avg + Std*
Random occupancy patterns [7]	86.5	—	—
Actionlet ensemble [8]	88.2	—	—
HON4D + D$_{disc}$ [9]	88.89	—	82.15 + 4.18
Spatial and temporal part sets [10]	90.22	—	—
HOPC of 3D pointclouds [11]	92.39	74.36	86.49 + 2.28
Points in a Lie group [12]	89.48	—	—
Histograms of action poses + DTW [13]	90.56	—	—
Dynemes and forward differences [14]	91.94	—	—
Part-based feature vector [15]	95.56	74.39	87.05 + 3.75
Ours	96.98	83.77	91.41 + 4.11

As shown in Table 1, we present the best result, the worst result and the average result compare with the state-of-the-art. Several studies have already been conducted on the MSRAction3D dataset. Our experimental results are presented in the last line of the table. Obviously, our method outperforms other methods in terms of the experimental result.



	high arm wave	horizontal arm wave	hammer	hand catch	forward punch	high throw	draw x	draw tick	draw circle	hand clap	two hand wave	side-boxing	bend	forward kick	side kick	jogging	tennis swing	tennis serve	golf swing	pick up & throw
high arm wave	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
horizontal arm wave	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
hammer	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
hand catch	0.05	0.00	0.00	0.90	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
forward punch	0.00	0.00	0.00	0.00	0.96	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
high throw	0.00	0.00	0.00	0.00	0.03	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
draw x	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
draw tick	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
draw circle	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
hand clap	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.94	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00
two hand wave	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
side-boxing	0.00	0.00	0.00	0.06	0.00	0.03	0.00	0.00	0.00	0.00	0.00	0.91	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
bend	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.05	0.90	0.00	0.00	0.00	0.00	0.00	0.00	0.00
forward kick	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
side kick	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
jogging	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
tennis swing	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
tennis serve	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.93	0.00	0.07
golf swing	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
pick up & throw	0.00	0.00	0.06	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.75

FIGURE 11. The confusion matrix of the best result

The confusion matrix of the best result is displayed in Figure 11. In the confusion matrix, the vertical coordinate (Y) represents the true label of an action sequence and the horizontal coordinate (X) represents the recognition result. The value at the (x, y) coordinate of the matrix represents the ratio of action y recognized as action x. From the confusion matrix of the best result in Figure 11, we can see that 12 out of 20 actions achieve 100% accuracy. Considering the noise of skeleton collected by

Kinect and some actions are very similar, the other actions did not reach 100% recognition rate is also normal.

6 Conclusion

In this article, we use a thought of hierarchy to design a two-level hierarchical recognition model based on MSRAction3D dataset. At the first level, we use the barycenters of five parts of the human body, and calculating the centralizing trend and discrete limit over time, that is, the mean and variance as features. Then we use SVM to classify all categories into seven categories. At the second level, we extract the relative position and relative velocity over time. Then we use the HMM to reclassify seven categories. Our method performed excellent recognition effectiveness based on the public dataset which extracted from the Kinect. It shows that the hierarchical recognition framework based on human skeleton is a feasible and effective method for HAR. Hierarchical identification strategy is very fit for the internal mechanism of human action. Around this work will continue in our next step.

Acknowledgements

The authors would like to thank the supports by the National Nature Science Foundation of China under Grant No. 11471093, National Science and technology of China No. 2014BAH13F02, Natural Science Research Funds of Education Department of Anhui Province (KJ2014A142) and Anhui Province Key Laboratory of Affective Computing & Advanced Intelligent Machine (ACAIM160102).

References

[1] Zhang J, Li W, Ogunbona P O, et al. RGB-D-based action recognition datasets: A survey[J]. Pattern Recognition, 2016, 60: 86-105.

[2] Su B, Jiang J, Tang Q, et al. Human periodic activity recognition based on functional features[C]. SIGGRAPH ASIA 2016, Symposium on Education. ACM, 2016: 6.

[3] Zhang Z. Microsoft Kinect sensor and its effect[J]. IEEE Multimedia, 2012, 19(2): 4-10.

[4] Papadopoulos G T, Axenopoulos A, Daras P. Real-time skeleton-tracking-based human action recognition using kinect data[C]. Proceedings of the International Conference on Multimedia Modeling, 2014: 473-483.

[5] Chen L, Wei H, Ferryman J. A survey of human motion analysis using depth imagery[J]. Pattern Recognition Letters, 2013, 34(15): 1995-2006.

[6] Li W, Zhang Z, Liu Z. Action recognition based on a bag of 3d points[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR)-Workshops, 2010: 9-14.

[7] Wang J, Liu Z, Chorowski J, et al. Robust 3d action recognition with random occupancy patterns[C]. Proceedings of the 12th European Conference on Computer Vision(ECCV), 2012: 872-885.

[8] Wang J, Liu Z, Wu Y, et al. Mining actionlet ensemble for action recognition with depth cameras[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2012: 1290-1297.

[9] Oreifej O, Liu Z. HON4D: Histogram of oriented 4d normals for activity recognition from depth sequences[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2013: 716-723.

[10] Wang C, Wang Y, Yuille A L. An approach to pose-based action recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2013: 915-922.

[11] Rahmani H, Mahmood A, Huynh D Q, et al. HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition[C]. Proceedings of the European Conference on Computer Vision(ECCV), 2014: 742-757.

[12] Vemulapalli R, Arrate F, Chellappa R. Human action recognition by representing 3d skeletons as points in a lie group[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2014: 588-595.

[13] Barnachon M, Bouakaz S, Boufama B, et al. Ongoing human action recognition with motion capture[J]. Pattern Recognition, 2014, 47(1): 238-247.

[14] Kapsouras I, Nikolaidis N. Action recognition on motion capture data using a dynemes and forward differences representation[J]. Journal of Visual Communication and Image Representation, 2014, 25(6): 1432-1445.

[15] Chen H, Wang G, Xue J H, et al. A novel hierarchical framework for human action recognition[J]. Pattern Recognition, 2016, 55: 148-159.