Abstract—Action recognition is a critical task for social robots to meaningfully engage with their environment. 3-D human skeleton-based action recognition has been an attractive research area in recent years. Although the existing approaches are good at action recognition, it is a great challenge to recognize a group of actions in an activity scene. To tackle this problem, at first, we partition the scene into several primitive actions (PAs)-based upon motion attention mechanism. Then, the PAs are described by the trajectory vectors of the corresponding joints. After that, motivated by text classification based on word embedding, we employ a convolutional neural network (CNN) to recognize activity scenes by considering motion of joints as “word” of activity. The experimental results on the dataset of human activity scenes show the efficiency of the proposed approach.

Index Terms—Convolutional neural network (CNN), primitive actions (PAs), scene recognition.

I. INTRODUCTION

ACTION recognition is an active topic in computer vision which aims at marking video frames with proper action labels [1], [2]. It is now widely applied in human–robot interaction, intelligent perception of social robots, visual surveillance, and video retrieval. Action recognition is usually composed of three major steps, that is: 1) action segmentation; 2) representation; and 3) classification [3]. Existing works tend to split the actions in the video frame by frame, or by several frames, which results in a large amount of features.

We know that action recognition can be seen as an intermediate stage that can provide a more complex system of interpreting, such as human behavior analysis and activity scene identification [4], [5]. In fact, recognizing the scene of human activity consisting of a group of actions is still a highly challenging task.

In this paper, we introduce an activity scene description, construction, and recognition method based on a 3-D skeleton sequence to tackle this problem. We first partition the scene of human activity into different primitive actions (PAs) based upon the kinematics of joint points. Then, an action descriptor, which is able to distinguish the difference between PAs, is proposed to describe these activities. Finally, motivated by word embedding-based text classification, we consider the motion of the joints as the “word” of the activity, and exploit a convolutional neural network (CNN) to recognize the activity scenes.

The key contributions of this paper can be summarized as follows.

1) We partition the scene of human activity into PAs according to the latest research result that the speed information and temporal cues are the two most important factors in tracking a moving object. Through partitioning the scene, the features are condensed remarkably.
2) The features extracted from skeleton sequence and used to describe the PAs are composed of both spatial and temporal information, which helps to effectively improve the recognition accuracy.
3) By regarding PAs as the words of human activity, scene recognition is implemented using a CNN that is very efficient in word embedding-based topic classification.

This paper is organized as follows. A review of related work is presented in Section II. Section III describes the proposed activity scene description, construction, and recognition method, followed by the experimental evaluation in Section IV. Finally, we conclude this paper in Section V.

II. RELATED WORK

In this section, we will introduce the state-of-the-art work involved in action recognition, mainly including the two categories of approaches proposed in the literature in recent years: one was based on the 2-D video stream and the other was based upon the 3-D skeleton joint sequence. In the first category, hand-crafted features, such as the histogram of gradient (HOG), histogram of optical flow (HOF), motion boundary histogram (MBH), and point trajectories, which have been popularly used to represent human actions, were extracted from a raw video stream for action classification [6], [7]. In order to overcome these features’ limitations of lacking semantics information and poor discriminative capacity, deep learning methods were proposed to automatically learn the semantic representation from raw video [8]–[12]. Recently,
Varol et al. [13] proposed a long-term temporal neural network with increased temporal extent to further improve the accuracy of action recognition.

With the rapid development of 3-D human pose estimation technology and the wide usage of the depth camera, 3-D human skeleton joints are able to be extracted in real time [14], [15]. Recently, the researchers pay much more attention on the methods based on 3-D skeleton joints. In this category, spatial, temporal, or trajectory information were extracted from skeleton joints to represent the human action [16]–[21]. The combination of distinct features showed the advantages in improving the recognition accuracy in [3] and [22]. Especially, a new framework for action recognition was proposed in [3] by means of the learning feature combination from the skeleton sequences. The impressive results had been illustrated in their literature with an average of more than 90% on accuracy. However, several limitations exist in these methods: 1) they just addressed the specific human actions which are composed of the short, simple, and well-defined sequences of motion and 2) the features extracted from skeleton joints are generated frame by frame and, therefore, have higher computational complexity.

III. OBJECT ACTIVITY SCENE DESCRIPTION, CONSTRUCTION, AND RECOGNITION

A. Pipeline

The pipeline of the proposed method is illustrated in Fig. 1. In the first stage, the human skeleton joints are divided into seven parts according to the kinetic analysis of the moving joints in the activity scenes. The detailed analysis is shown in Section III-B. Recent research has proved that in the process of visually tracking a moving object, humans are generally most sensitive to the speed and time of the moving object, that is, the speed and time are the two most important factors in object tracking. We call it motion attention mechanism. Under the guidance of this mechanism, we use the synthetic motion parameter of each group of the part as the indicator to partition the primitive motions separately. Then, the representative features are extracted to describe the partitioned motions. Since the spatial and temporal information are considered, these features help to improve the recognition accuracy. In the final stage, a CNN is exploited to recognize the scenes and output the resulting labels. We give the details of this part in Section III-D.

B. Object Activity Scene Construction

In this section, we introduce the details of how to construct the human actions. Intuitively, a human action consists of several primitive motions that can be regarded as the atomic movements used to express specific semantic behavior. Recently, an actionlet ensemble method was proposed to discover discriminative motions by using data mining from several hand-crafted features [22]. The limitation is that the motions of the joints are assumed to be independent. On the contrary, they are coupled to each other and are very complicated even if they are described separately.

1) Body Parts Division: We know that a human’s movement includes the motion of the entire body (if we regard human as a particle) and the changes in local relative posture (if we use a specific point on the body as a reference). Besides the motion introduced by entire body’s movement in action recognition, by constructing local coordinate systems (LCSs), we also emphasize the relative motion parameters of the joints under the assumption that the majority of actions can be recognized by the local posture.

a) Coordinate system: To properly describe the motion of the human body, two coordinate systems are constructed in this paper, that is, the global coordinate system (GCS) and LCS, as shown in Fig. 2(a). GCS is a right-handed coordinate system that places a camera at the origin with the positive z-axis extending in the direction in which the camera is pointed [16]. Three joints are chosen to construct LCS,
that is, the spin, the left, and right hip joints, as illustrated in Fig. 2(b), because they are more suitable to describe the motion of the human body as a whole. The foot drop of the connection of the left and right hip joints and the plain which passes through the spin is set to the origin of LCS (denoted by “root” joint), and the positive $x$-axis points to the left hip joint, and $y$-axis starts from the origin to the spin joint. The right-handed rule is also true in LCS.

b) Construction of the division pattern of the human joints: From the kinematic perspective, we know that the displacement, the speed, and the acceleration are three important indicators of motion. Actually, human actions are extremely difficult to describe, because the motions of human joints have strong arbitrariness and some of them are also highly related. In order to efficiently describe the motion of the joint, a feasible solution is to construct a human joints division pattern, that is, dividing the human skeleton joints into several groups, and then describing them individually.

If we examine each joint separately, the dimension of the feature vector will be large and the computational complexity will easily increase during the recognition phase. The criteria of dividing human joints rely on two aspects: the joints with the same motion pattern should be correctly divided into the same group, while significantly reducing the dimension of the feature vector.

Du et al. [23] divided all of the human skeleton joints into five parts, that is, two arms, two legs, and one trunk, as shown in Fig. 3(a), but the limitation is that the joints with the different motion pattern do not separate properly which will lead to the degradation of recognition accuracy. For example, we infer that the 13 lhumerus joint and the 14 lhand joint may have a similar pattern, because the forearm is always driven by the upper arm to produce the local motion. On the contrary, the 12 lclavicle joint is highly correlated to the upper torso and is more suitable to reflect the movement of the entire body.
In order to prove our inference, we define a function to measure the similarity of the motions of different joints. At first, we give some definitions. It is assumed that the joint location \((x, y, z)\) of the human skeleton in the 3-D coordinate can be handled directly. The speed of the joint is defined as follows:

\[
v_{f, j, i} = \frac{s_{f, j, i} - s_{f-1, j, i}}{\Delta t}, \quad i \in \{1, 2, 3\} \quad (1)
\]

where \(s_{f, j, i}\) is the location of the \(j\)th joint in the \(i\)th coordinate axis and \(f\)th frame, and \(\Delta t\) is the time interval between two frames.

Then, the synthetic speed of the joints in three directions is denoted by

\[
v_f = \sqrt{\sum_{i=1}^{3} (v_{f, i})^2} \quad (2)
\]

and the corresponding synthetic acceleration of the \(j\)th joint can be written as

\[
d_f = \frac{v_f - v_{f-1}}{\Delta t}. \quad (3)
\]

The Euclidean distance between two joints’ synthetic acceleration is employed to measure their motion similarity, and the function is given by

\[
D_{j, k} = \text{dist}(a_j, a_k) \quad (4)
\]

where \(a_j\) and \(a_k\) are the sequences of synthetic acceleration for joint \(j\) and \(k\), respectively.

We calculate the motion similarity of several selected joints using the function (4), and the results are shown in Table I.

| Selected Joints | \(D_{12,13}\) | \(D_{12,15}\) | \(D_{15,16}\) | \(D_{22,3}\) | \(D_{22,5}\) | \(D_{5,6}\) |
|-----------------|--------------|--------------|--------------|-------------|-------------|-------------|
| Motion Similarity | 1.03E+04     | 4.69E+03     | 8.64E+03     | 9.34E+03    | 4.11E+03    | 9.31E+03    |

From Table I, we can see that \(D_{12,15}\) is much smaller than \(D_{12,13}\) and \(D_{15,16}\), and the same results can be observed in hip joints (2 and 5) and leg joints (3 and 6).

To show the similarity more intuitively, we draw the synthetic speed curves of some joints, as illustrated in Fig. 4. From Fig. 4, we can see that the motions of the 13 humerus joint and the 14 lhand joint have higher similarity than the 12 lclavicle joint on time sequence, but the 15 rclavicle joint is highly related to the 12 lclavicle joint. The similar pattern can be noticed in the motion of other joints, which uncovers the fact that the function defined in (4) represents the inherent characteristic of motion patterns of these joints. According to the above observation, seven parts are divided in this paper which are listed in Table II and are shown in Fig. 3(b), and their corresponding pivot joints (or rotation center) are also illustrated (green solid points). The torso joints are usually used to reflect the global motion of humans, such as moving forward, backward, left and right, and jumping up and down. On the contrary, the limbs joints are more suitable to represent local motions. In this paper, the head motion is ignored.
2) **Primitive Motion Partition:** It is a challenge to describe a human’s action, because of the strong dependency of the velocity and position of human joints. Suppose we consider the recognition problem as a function approximation from a mathematical perspective (which is reasonable when we use the neural network to tackle this problem), and motivated by defining primitive functions in the parameters neural networks in [24] and [25], we attempt to partition the scene of human activity into the PAs.

By observing human actions from the frame level, we find that each human part (as listed in Table II) usually has a standstill between two different motions. Two successive standstills can be used to represent the start and the end of a primitive motion. Therefore, in this paper, we consider using the standstill as the indicator of the interval of two primitive motions. The end joints in seven human parts are chosen to partition the motions, because these joints in each part are highly related and have similar motion patterns as explained in Section III-B1.

The synthetic speed of the joints is employed to partition the actions, and it has been defined in (4). We notice that the joint positions should be preprocessed to eliminate the negative effect of wild values; meanwhile, insignificant motions should be also filtered to simplify the action recognition. The preprocessed synthetic speed is

\[
\bar{v}_j^f = \begin{cases} 
  v_j^f, & \text{if } v_j^f \geq v_\tau \\
  0, & \text{otherwise}
\end{cases}
\]

(5)

where \(v_\tau\) is the threshold which is used to suppress the negligible motions.

The recent research of the psychophysicists from MIT revealed that in the process of visually tracking moving objects, humans rely on both speed information and temporal information [26]. This means that humans are generally more sensitive to the speed and time of moving objects, which can be considered as a motion attention mechanism. Therefore, the joints with higher moving speed and longer displacement (represented as the product of speed and time) will receive more attention when partitioning the PAs. Based on the above consideration, in the time sequence of moving joints, the intervals at which the synthetic velocity of joints have a larger value and a greater width will be considered as the primitive motions.

Since each human part has two joints, the intervals of the two joints are combined, which is denoted as follows:

\[
M_{p,q} = \bigcup_n \left( S_n | \bar{v}_j^f \neq 0 \right), \quad n = \{1, 2\}
\]

(6)

where \(n\) is the number of joints in the part, \(M_{p,q}\) is the \(q\)th combined interval of the \(p\)th human part, and \(S_n\) is the interval selected to represent the primitive motion. The motion in each combined interval is partitioned as a primitive motion of this part. The number of the primitive motions in the activity scenes can be regarded as a super-parameter. The partition results are shown in Fig. 5, in which only three sets of end joints are drawn as an example. The adjacent primitive motions are distinguished by different colored dotted lines.

C. **Primitive Motion Descriptor**

After partitioning the primitive motions of these joints, it is extremely important to describe them in the following recognition stage. In this section, we present a PA descriptor consisting of two parts: 1) global motion descriptor and 2) local motion descriptor.

1) **Global Motion Descriptor:** Global motion is described by three joints, that is, the root, lhip, and rhip joints, in the lower torso part. To avoid redundancy, we just consider the lhip and rhip joints. The global features are extracted from the trajectory of these two relevant joints. The joint trajectory...
is a sequence of points in 3-D space that describes the motion path of the specified joint, and it can be represented by a function

\[ T_j^q = f(b_j^q, m_j^q, e_j^q) \]  \hspace{1cm} (7)

where \( T_j^q \) is the trajectory of the \( j \)th joint in the \( q \)th primitive motion, and \( b_j^q \) and \( e_j^q \) are the start and end displacement vectors related to the origin of GCS which are denoted as

\[ b_j^q = s_j^0 - 0, \quad e_j^q = s_j^f - 0 \]  \hspace{1cm} (8)

respectively, and \( m_j^q \in \mathbb{R}^3 \) is the intermediate points in the trajectory, as shown in Fig. 6(a). In this paper, we choose five uniformly distributed intermediate points from the trajectory. For simplicity, we drop the superscript in the following section.

2) **Local Motion Descriptor:** The description of the local motion has some difference with the global motion. For one thing, the different parts of the human body should be described separately. That means that the local motion features are composed of a total of seven parts’ feature. For another, we are only interested in the relative motion of human parts in the local environment. The relative trajectory can be denoted by

\[ T_j^q = f(b_j^q, m_j^q, e_j^q). \]  \hspace{1cm} (9)

The difference between (7) and (9) is that the parameters of the \( j \)th joint in (9) are all related to the corresponding pivot joint. We take the right arm as an example in Fig. 6(b), and its features are composed of the rhumerus’s trajectory vector \( rT_{16} \) related to the rclavicle, and the rhand’s trajectory vector \( rT_{17} \) related to the rhumerus. The strategy used in the selection of the intermediate points is similar to the global motion description.

a) **Extension of local features:** When we obtain the features of each part according to (9), the motion of end joints in this part is clarified. However, in a natural sense, different parts of the human body are bonded rather than isolated when they are combined to represent a specific activity. In this paper, we construct this relationship between the joints in different parts through extending the local features of each part are not only in relation to the pivot joint but also in relation to the joints in other parts. After that, the local features of the primitive motions for each part are concatenated, and the features for global and local motions are overlaid according to chronological order.

From the features used to describe the primitive motion, we can find that both spatial and temporal information are contained. The temporal relationship is reflected in the primitive motion vectors of each part which are arranged in chronological order, while the extended local features maintain a spatial bonding between the different parts.

3) **Feature Normalization:** It is worth noting that the local features extracted from human motion depicted in Section III-C2 are completely relative displacement vectors. This means that they have latent drawbacks when they encounter differences in the human body. For example, the features are significantly different when they are extracted from the same action but conducted by an adult and a kid, because the results are affected by the length of their bodies. To eliminate the impact of the difference of the human body, the normalization step is followed by motion description. For each displacement vector, the normalized feature is generated by dividing the norm of the corresponding vector.

D. **Scene Recognition Based on the Convolutional Neural Network**

We know that before language was invented, action was the most important manner for humans to communicate with each other. Human beings deliver specific semantics through a series of actions with spatial and temporal relationships. It is reasonable to regard these actions as body language compared to words in natural language. Correspondingly, a group of actions, that is, an activity scene, used to deliver the semantic can be considered as a “paragraph.” The task of recognizing the scenes of human activity from the video is similar to classifying a paragraph of words into different topics, that is, text classification.

Combining word embedding with CNN is an effective approach popularly used by the authors in the field of text classification. Motivated by text classification, we employ CNN to recognize activity scenes by considering the actions of joint as the words of the activity. CNN has been employed in human
action recognition in recent articles because of its promising performance in classification. However, the majority of existing approaches were conducted on raw image sequence. In this paper, the proposed method is based on the 3-D skeleton joints and the PA description and representation introduced in the above sections. The CNN model is shown in Fig. 7. As can be seen from Fig. 7, the model consists of four layers of a deep neural network in the order of convolutional layer, max-pooling layer, fully connected layer, and softmax layer. The convolutional layer with multiple filter widths has the capability to learn spatial and temporal relationships of the PAs from the features extracted.

Compared with the deep neural networks used in the field of computer vision, such as image classification [27]; object detection [28]; scene classification [29], [30]; and even change detection [31], the network used for the activity recognition in this paper is relatively simple. However, the sophisticated hand-crafted features can be extracted from the 3-D skeleton based on humans’ prior knowledge. With these features in hand, the network can very effectively distinguish different activities, which will be shown in Section IV.

IV. EXPERIMENTAL RESULTS

It is noticed that there are plenty of datasets available for the action recognition methods based on the 3-D skeleton. Existing algorithms usually evaluate their performance on three popular datasets: 1) MSR-Action3D [32]; 2) UTKinect-Action3D [16]; and 3) Florence 3-D Actions [33], and their average accuracies are up to 98% [3], [23], [34], [35]. However, these databases are relatively simple because the actions only involve the movement of a single component (e.g., left arm and right arm) or a combination of a small number of components. In our experiments, we consider a more challenging activity: scene recognition task.

A. Dataset

1) H36m Dataset: The H36m dataset [36] contains 3.6 million 3-D human skeleton and corresponding frames collected by four digital cameras, one time-of-flight sensor, and ten motion cameras, with 11 professional actors (6 male and 5 female) and 17 scenes. We use a publicly opened subset with 422,055 frames which are downsampling from the original videos (50 frames/s) (in order to reduce the correlation of the consecutive frames). The subset contains 15 scenes associated with seven subjects whose ground-truth 3-D skeleton are provided. The 15 activity scenes include: 1) discussion; 2) direction; 3) eating; 4) greeting; 5) posing; 6) purchases; 7) sitting; 8) sitting down; 9) smoking; 10) waiting; 11) walking; 12) walk together; 13) walk dog; 14) taking a photograph; and 15) talking on the phone. Each scene contains a series of actions to express a specific semantic activity that is closer to the natural interaction scenarios. This dataset is much more challenging because: 1) the length of the sequences is quite long on average and varies greatly (from 990 to 6340 frames); 2) the diversity within the same category is very large, for example, for “posing;” different people complete the action according to their own understanding; and 3) the dataset contains some confusing actions, such as walking, walk together, and walk dog, as well as sitting and sitting down.

2) HDM05 Dataset: The HDM05 dataset contains more than three hours of systematically recorded and well-documented motion capture data with 22 activity scenes. Most of the motion sequences have been performed several times by all five actors according to the guidelines of the fixed script. We select 13 activity scenes from the original dataset, and most of the actors perform no less than three repetitions. The selected activity scenes include walking, locomotion on the spot, locomotion, locomotion with weights, table and floor, shelf (while walking), shelf (while standing), dancing, kicking and punching, throwing, rotating arms, workout, and clapping
and waving. We downsample the frame rate from 120 frames/s to 24 frames/s. Then, the length of the motion sequences varies from 318 to 2738, and the average length is about 1195 frames.

B. Dataset Augmentation

Although deep neural networks have made great achievements in a wide aspect of intelligent tasks, such as text and image classification, natural language processing, and action recognition, one of the main challenges in using neural networks in 3-D action recognition is that the number of available training samples is relatively small, often leading to overfitting. Intuitively, seeking the way to augment the dataset has a higher priority to eliminate the drawback of overfitting. However, it is difficult to extend the existing datasets due to the inevitable bias introduction during collecting, processing, and validating new data samples.

We assume that the same semantics can be conveyed when people perform the same action using left limbs and right counterparts. This is always true in real human–human interactions and human–robot interactions. For example, some people like to drink water with their left arms, but others prefer to using their right arms. Resorting to the spatial properties of 3-D skeleton joints, we are able to construct the mirror model of human actions. Based on the LCS established in Section III, we use the $yoz$ plane as the symmetry plane to flip the coordinates of the human joints to obtain the new coordinates of each joint, thereby doubling the original sample size.

C. Parameter Setting

Through augmenting the original H36m dataset, the total of 1680 scene samples can be obtained. We split them into three parts: the two subjects (480 scenes) are used for the validation set and the testing set, respectively, and the remaining five subjects (1200 scenes) are employed for the training set. We split the HDM05 dataset using a strategy similar to the H36m dataset. The difference is that there are three subjects’ activity samples in the training set of the HDM05 dataset.

To determine the total number of aforementioned PAs in Section III-C2, we calculate the PAs extracted from the dataset, as shown in Fig. 8. We can see that the number of PAs is between 0 and 40; therefore, the maximum amount of PAs is set to 30 in order to balance the efficiency and compactness of the features.

To reveal the sensitiveness of performance on parameters, we illustrate three groups of experimental results with different numbers of filters in a convolutional layer and different learning rate. The experiments on parameter sensitiveness are conducted on the H36m dataset. In the first experiment, we fix the learning rate to $10^{-4}$ and set the filters number to 768. Then, we fix the number of filters to 1024, and conduct the remaining two experiments with the learning rates of $10^{-3}$ and $10^{-4}$, respectively. The number of neurons for the fully connected layer is 256 and the implementation is based upon Tensorflow.

As we know, many distinguished action recognition approaches have been proposed in recent years. It is difficult to evaluate all of their performances in recognizing activity scenes in the H36m dataset. In this paper, only five representative methods are chosen for comparison, as shown in Table III. The average frame of the H36m dataset is about 500, so the fixed feature length $T$ for the spatial–temporal RNN [21] is set to 500, and it is set to 1200 for the HDM05 dataset correspondingly. The other parameters of these methods coincide with the original articles. Especially, for a six-layer RNN, a deep LSTM model is used and the dimension of hidden neurons is set to 256. All simulations are run on an Intel Core i7-4790 CPU with 32-GB memory and Nvidia 1080 Ti GPU.

D. Results Comparison and Analysis

1) H36m Dataset: The experimental results conducted on the H36m dataset are shown in Table III. From Table III, we can see that the accuracy of the proposed method is 6.64% higher than the best result based on the vector of locally aggregated descriptors (VLADs) with metric learning [3], and our method also outperforms LSTM-based spatial–temporal RNN (7.72%) and six layers RNN (33.9%). Since the trust gate is introduced in the LSTM, the spatial–temporal RNN has much better improvement (26.18%) than the traditional six-layer RNN.

It can also be observed that the performance of the proposed method is more sensitive to the learning rate than the number of filters in the model. The confusion matrix of our method is shown in Fig. 9. Several scenes are recognized without
any mistake, and most of the accuracy is high. It is noticed that the classification of two scenes (“photograph” and “smoking”) is completely wrong. To find out the reason, we take the photograph scene as an example, and compare the left arm’s synthetic speeds (which are used to partition the PAs) in the photograph scene in the training and testing sets and “direction” scene in the training set, as shown in Fig. 10. From Fig. 10, we can see that the photograph scene in the testing sets contains more detailed actions. On the contrary, there are fewer actions in the training set. Therefore, in this situation, the classifier is more likely to recognize it as the direction scene than the photograph scene. The similar phenomenon can be seen in the other misclassified activity scenes. This is the reason why the H36m dataset is more challenging than the other datasets.

2) HDM05 Dataset: For the HDM05 dataset, the comparable experimental results are shown in Table IV. As can be seen from Table IV, the accuracy of our method is 92.31%, which is the best result on this dataset, and it is about 3.77% higher than the nearest competitor VLAD with metric learning. In this case, all of the methods achieve a higher score than the results tested on the H36m dataset, which can be also revealed in the confusion matrix, as shown in Fig. 11. Eleven out of thirteen activity scenes are classified without any mistake. The reason is that all five actors perform according to the guidelines of the fixed script in the HDM05 dataset, so the actors are less flexible and arbitrary than the H36m data set. The worst result is obtained on “rotate arms,” and it is classified as “throwing” with the probability of 66.7%. By comparing the original videos on these two activities, we can notice that there are several rotating arm actions in the throwing activity. Even humans do not easily recognize them correctly.

E. Impact of the Number of Intermediate Points

In the proposed primitive motion descriptor (Section III-C), the intermediate points are selected to represent the trajectory of the joints. There is no a priori knowledge of how to determine the number of intermediate points. In this section, we conduct a group of experiments on the H36m dataset to reveal the impact of this parameter, and the result is shown in Fig. 12. Fig. 12 illustrates the relationship between the number

| Method               | Accuracy |
|----------------------|----------|
| Lie Group + SVM      | 83.22%   |
| VLAD + Metric Learning | 88.54%   |
| Six Layers RNN       | 63.38%   |
| Spatial-temporal RNN | 72.44%   |
| Proposed CNN         | 92.31%   |

Table IV

Experimental Results on the HDM05 Dataset

Fig. 12. Graph illustrating the relationship between the number of intermediate points and accuracy.
of intermediate points and the recognition accuracy as well as the prediction time. We can see that the peak accuracy is obtained when five intermediate points are selected, and the accuracy drops dramatically when the number of points is greater than ten. Meanwhile, we can also notice that the prediction time increases as the number of intermediate points increases, because the more points that are chosen, the longer the feature length.

**F. Effect of the Global Motion Feature**

In Section III-C, the proposed motion descriptor includes both the global motion and the local motion. The global motion represents the movement of the entire body, while the local motion describes the relative motion between body parts. From Section III-C1, we can note that only three joints are used to describe the global motion, so we compare the recognition performance of whether the global motion features are combined in the local motion features. The experimental result is illustrated in Table V. As shown in Table V, as the feature length is increased by 6300, the descriptor combining the global motion and the local motion obtains better performance of recognition than only using the local motion features (with 2.92% improvement).

**G. Computation Time**

The computation time of action recognition usually includes feature extraction time, training time, and testing time. We make a comparison of the computation time of selected methods, and the results are shown in Table VI (the unit is second). From Table VI, we can see that the proposed method consumes a shorter total time than the other methods. It is noticed that although the average training time for VLAD is shorter, the time for each epoch is not equal because the dimension of the feature decreases while epochs increase.

**V. Conclusion**

In this paper, we proposed a novel object activity scene description, construction, and recognition method. Targeting the limitations of existing approaches in recognizing activity scenes, we proposed partitioning the scene into several PAs based upon the motion attention mechanism, and then well-defined features containing both spatial and temporal information are extracted to describe these PAs. Motivated by text classification, a CNN is employed for scene recognition by regarding the actions of the joint as words and the corresponding activity scene as a paragraph. Experimental results reveal that the steps of the proposed method are of great significance to improve the accuracy of scene recognition. By comparing the method with existing algorithms, we find that the proposed method outperforms them in terms of recognition accuracy and time complexity.

In our future work, we should carefully consider the fact that 3-D joint sequences are not available for some real applications, such as real-time human–robot interaction. In this situation, the performance of our method will be affected to some extent; however, human motion kinetic modeling as well as filtering methods can be used to eliminate this effect.

**ACKNOWLEDGMENT**

The authors would like to thank the reviewers and an associate editor for their constructive suggestions.

**REFERENCES**

[1] R. Poppe, “A survey on vision-based human action recognition,” Image Vis. Comput., vol. 28, no. 6, pp. 976–990, 2010.

[2] D. Weinland, R. Ronfard, and E. Boyer, “A survey of vision-based methods for action representation, segmentation and recognition,” Comput. Vis. Image Understand., vol. 115, no. 2, pp. 224–241, 2011.

[3] D. C. Lavizon, H. Tabia, and D. Picard, “Learning features combination for human action recognition from skeleton sequences,” Pattern Recognit. Lett., vol. 99, pp. 13–20, Nov. 2017.

[4] V. Magnanimo, M. Saveriano, S. Rossi, and D. Lee, “A Bayesian approach for task recognition and future human activity prediction,” in Proc. IEEE Int. Symp. Robot Human Interact. Commun., 2014, pp. 726–731.

[5] H. He, S. S. Ge, and Z. Zhang, “Visual attention prediction using saliency determination of scene understanding for social robots,” Int. J. Soc. Robot., vol. 3, no. 4, pp. 457–468, 2011.

[6] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2008, pp. 1–8.

[7] H. Wang and C. Schmid, “Action recognition with improved trajectories,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Sydney, NSW, Australia, 2013, pp. 3551–3558.
[8] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, “Convolutional learning of spatio-temporal features,” in Proc. Eur. Conf. Comput. Vis. (ECCV), 2010, pp. 140–153.
[9] S. Ji, W. Xu, M. Yang, and K. Yu, “3D convolutional neural networks for human action recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 221–231, Jan. 2013.
[10] A. Karpathy et al., “Large-scale video classification with convolutional neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2014, pp. 1725–1732.
[11] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in Proc. Adv. Neural Inf. Process. Syst., 2014, pp. 568–576.
[12] L. Wang, Y. Qiao, and X. Tang, “Action recognition with trajectory-pooled deep-convolutional descriptors,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, 2015, pp. 4305–4314.
[13] G. Varol, I. Laptev, and C. Schmid, “Long-term temporal convolutions for action recognition,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 6, pp. 1510–1517, Jun. 2018.
[14] J. Shotton et al., “Real-time human pose recognition in parts from single depth images,” Commun. ACM, vol. 56, no. 1, pp. 116–124, 2013.
[15] Y. Ma et al., “VNeck: Real-time 3D human pose estimation with a single RGB camera,” ACM Trans. Graph., vol. 36, no. 4, pp. 1–14, 2017.
[16] L. Xia, C.-C. Chen, and J. K. Aggarwal, “View invariant human action recognition using histograms of 3D joints,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2012, pp. 20–27.
[17] J. Luo, W. Wang, and H. Qi, “Group sparsity and geometry constrained dictionary learning for action recognition from depth maps,” in Proc. IEEE Conf. Comput. Vis. (ICCV), Sydney, NSW, Australia, 2013, pp. 1809–1816.
[18] R. Vemulapalli, F. Arrate, and R. Chellappa, “Human action recognition by representing 3D skeletons as points in a lie group,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Columbus, OH, USA, 2014, pp. 588–595.
[19] M. Devanarayan et al., “3D human action recognition by shape analysis of motion trajectories on Riemannian manifold,” IEEE Trans. Cybern., vol. 45, no. 7, pp. 1340–1352, Jul. 2015.
[20] R. Slama, H. Wannous, M. Daoudi, and A. Srivastava, “Accurate 3D action recognition using learning on the Grassmann manifold,” Pattern Recognit., vol. 48, no. 2, pp. 556–567, 2015.
[21] H. Wang and L. Wang, “Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Honolulu, HI, USA, 2017, pp. 499–508.
[22] J. Wang, Z. Liu, Y. Wu, and J. Yuan, “Mining actionlet ensemble for action recognition with depth cameras,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2012, pp. 1290–1297.
[23] Y. Du, W. Wang, and L. Wang, “Hierarchical recurrent neural network for skeleton based action recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, 2015, pp. 1110–1118.
[24] C. C. De Wit and S. S. Ge, “Adaptive friction compensation for systems with generalized velocity/position friction dependency,” in Proc. IEEE Conf. Decis. Control, vol. 3, San Diego, CA, USA, 1997, pp. 2465–2470.
[25] S. S. Ge, T. H. Lee, and C. J. Harris, Adaptive Neu,” Pattern Recognit. Control of Robotic Manipulators, vol. 19. Singapore: World Sci., 1998.
[26] C.-J. Chang and M. Jazayeri, “Integration of speed and time for estimating time to contact,” Proc. Nat. Acad. Sci. USA, vol. 115, no. 12, pp. 2879–2887, 2018.
[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS), vol. 1, 2012, pp. 1097–1105.
[28] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778.
[29] B. Zhou, A. L. Garcia, J. Xiao, A. Torralba, and A. Oliva, “Learning deep features for scene recognition using places database,” in Proc. Int. Conf. Neural Inf. Process. Syst. (NIPS), vol. 1, 2014, pp. 487–495.
[30] Q. Wang, S. Liu, J. Chanussot, and X. Li, “Scene classification with recurrent attention of VHR remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 2, pp. 155–167, Feb. 2019.
[31] Q. Wang, Z. Yuan, Q. Du, and X. Li, “GETNET: A general end-to-end 2-D CNN framework for hyperspectral image change detection,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 1, pp. 3–13, Jan. 2019.
[32] W. Li, Z. Zhang, and Z. Liu, “Action recognition based on a bag of 3D points,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2010, pp. 9–14.
[33] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, and P. Pala, “Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2013, pp. 479–485.
[34] O. Oreifej and Z. Liu, “HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Portland, OR, USA, 2013, pp. 716–723.
[35] X. Yang and Y. Tian, “Super normal vector for action recognition using depth sequences,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2014, pp. 804–811.
[36] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, “Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 36, no. 7, pp. 1325–1339, Jul. 2014.

Hui Feng received the B.Sc. and Ph.D. degrees in computer science from Huazhong University of Science and Technology, Wuhan, China, in 2007 and 2012, respectively. He is currently an Associate Professor with the School of Transportation, Wuhan University of Technology, Wuhan, China. He has published over 20 papers in conferences and journals, such as the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, ACM TOMM, Applied Soft Computing, and Multimedia Tools and Applications.

Shanshan Wang received the M.S. degree in computer science from Huazhong University of Science and Technology, Wuhan, China, in 2011. She is currently a Visiting Scholar with the Social Robotics Laboratory, Department of Electrical and Computer Engineering, National University of Singapore, Singapore. Her current research interests include computer vision, action recognition, and artificial intelligence.

Haixiang Xu received the B.Sc. degree in ocean engineering from Wuhan University of Technology, Wuhan, China, in 2001, and the Ph.D. degree in telecommunications from Huazhong University of Science and Technology, Wuhan, in 2005. He is currently a Professor with the School of Transportation, Wuhan University of Technology. He has published over 50 papers in conferences and journals. His current research interests include dynamic positioning, pattern recognition, and artificial intelligence.

Shuzhi Sam Ge (S’90–M’92–SM’99–F’06) received the B.Sc. degree in automatic control from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 1986 and the Ph.D. degree in mechanical and electrical engineering from the Imperial College of Science, Technology and Medicine, University of London, London, U.K., in 1993. He is the Founding Director of the Social Robotics Laboratory, Interactive Digital Media Institute, National University of Singapore, Singapore, where he is a Professor with the Department of Electrical and Computer Engineering. He has authored or co-authored six books and over 300 international journal and conference papers. His current research interests include autonomous robotics, intelligence control, computer vision, action recognition, and machine learning.

He is the Editor-in-Chief of the International Journal of Social Robotics. He has served as an Associate Editor for a number of flagship journals.