Research Article

Multiperson Interactive Activity Recognition Based on Interaction Relation Model

Hongbin Tu 1, 2, Renyu Xu 1, Rui Chi 1, and Yuanyuan Peng 1

1 School of Electrical and Automation Engineering, East China Jiaotong University, Nanchang 330013, China
2 Changsha Information Research Institute, Hunan Great Wall Technology Information Company Limited, Changsha 410000, China

Correspondence should be addressed to Hongbin Tu; tuhongbin310@163.com

Received 16 February 2021; Revised 30 June 2021; Accepted 30 July 2021; Published 12 August 2021

Copyright © 2021 Hongbin Tu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Multiperson activity recognition is a pivotal branch as well as a challenging topic of human action recognition research. This paper adopts a hybrid learning model to the spatio-temporal relationship and occlusion relationship among multiple people. Initially, this paper builds up an active multiperson interaction relationship estimation framework model to capture interpersonal spatio-temporal relation. This model incorporates the interaction relationship estimation framework with the multiperson relationship network. On this ground, it automatically learns from the human-computer interaction dataset in an end-to-end manner and performs reasoning with standard matrix operations. Secondly, this paper proposed an adaptive occlusion state behavior recognition method derived from the semantic knowledge model to unravel the concern of occlusion and self-occlusion in human action recognition. Then, Petri Nets are used to recognize multiperson interactive actions. This model has been through extensive experiments on the TV interaction dataset, Vlog dataset, AVA dataset, and MLB-YouTube dataset, experimental results have proved that the recognition performance of this model is superior than the other available models. This paper prospects and summarizes the estimation framework of the interaction relationship and occlusion semantic-knowledge relationship. Experimental results suggest that the proposed method in the paper could capture the discriminative relation information for multiperson interactive activity recognition, which further validates the efficiency of the hybrid learning model.

1. Introduction

Human interactive activity involves social behaviors and interactive behaviors. The former refers to individual behaviors, yet it takes other individuals’ behaviors into account; the latter refers to group behaviors for a shared goal. To recognize different behaviors, collective intelligence is thus built, aiming to jointly reason multiple individuals’ behaviors. On this ground, the paper proposed a method to recognize group behaviors which allow us to locate and narrate the interactive and collective actions of each individual in the context. The above perception of social context can be applied to sports analytics, social behavior recognition, and surveillance.

Recent methods for multiperson interactive behavior recognition take a sequential approach. Different from action recognition [1–8], which focuses on recognition of individual action, multiple behavior recognition settles down to figuring out the multiperson interactive behavior in a public area scenario, which can be applied to practical applications such as visual surveillance, intelligent robotics, and sport event analysis. Multiple activity involves more than one person and can scarcely be identified if only considering the action of a single person. For example, the action standing (performed by a single subject) is in fact compatible with multiperson interactive behavior, as shown in Figures 1–4. Hence, it is important to take the interactions overhead the related group into consideration for multiperson interactive activity modelling.

Dating back to a decade ago, interactive activity recognition modelling started with developing the model each class of collective activity by an interaction matrix [9] and
employed systematically organized skeleton features enhanced with directional features in dealing with interactive action recognition and real-time detection tasks [10]. An-An et al. [11] proposed visual feature learning which is incorporated into the multitask learning framework, with the Frobenius-norm regularization term and the sparse
constraint term, for joint task modelling and task relatedness-induced feature learning. Zhu et al. [12] proposed a photonic switched optically connected memory system architecture for deep learning models. However, these descriptors are not able to tell the exact relationship between people in a group. Xu et al. [13] proposed technology of novel group activity recognition which is proposed based on multimodal relation representation with temporal-spatial attention. Kim et al. [14] proposed a set of predictive features interbehavior relation based on spatial, temporal, transitional, and environmental contexts. Liu et al. [15] proposed a deep fully connected relation model to learn the interactions between people. Lu et al. [16] proposed a graph attention interaction model (GAIM) embedded with the graph attention block (GAB) to explicitly and adaptively infer unbalanced interaction relations at personal and group levels.

In these previous studies, features are first extracted for each person, and then, the interactions between each pair of features are explicitly modelled. A key limitation of most existing models lies in the separation of human detection and activity modelling. Existing models weigh more on group behavior or collective behavior modelling. As for human detection, they directly adopt bounding boxes output by the third-party pedestrian detectors. Unfortunately, this design completely decouples the detector and collective activity recognizer and abandons the inherent collaboration between these two modules. The outliers who perform individual activities semantically irrelevant to the main group or the missing active participant information are detrimental to the collective activity recognition (misclassified as crossing and walking, respectively), thus affecting group activity recognition. The lack of collaboration between the detector and recognizer brought by the separate learning/processing in previous methods also leads to heavy computation when reasoning the collective activities. This is because forward propagation through the neural network
backbone that aims at basic feature extraction is executed twice for the detection and activity recognition tasks, respectively.

We have also observed that the current trend [17–21] of tackling the problem of multiperson interactive activity recognition is to develop a model or framework with increasing complexity to jointly learn more subtasks simultaneously (e.g., detection, tracking, pose estimation, appearance modelling, and interaction).

Although approaches aforementioned seems to be reasonable, it has some limitations. Firstly, most of the advanced detection approaches do not involve joint optimization to process multiple objects, but rather rely heavily on the heuristic postprocessing approach. Greedy nonoptimal decisions are thus common to be seen. Secondly, feature extraction of each object ignores a large amount of context and interaction, which is believed to be productive information for reasoning multipersonal interactions because the position and action of some human objects with interactive relationships can be highly correlated. Thirdly, separating the tracking and the detecting suggest the loss of positioning features, and the utilization of these features will make a more productive recognition model. Finally, the sequential approach cannot prevail in a multiperson scene as it needs to be run multiple times on a image.

In short, previous approaches showcase some advantages (e.g., a good joint learning framework can process multiple tasks simultaneously), but it also reveals noticeable limits. For instance, (1) the sophisticated model which can be difficult to be optimized, (2) each subtask has not been fully explored and studied in depth, and (3) the core part and the problem-solving mechanism of these models remain underexplained, and it is difficult to guide further research studies.

The approach suggested in our research attempts to overcome the above limitations and address specific problems. Inspired by the recent multipersonal interaction recognition, we propose the multiperson interactive relation Graph (MRG) to simultaneously capture the interpersonal appearance and position relation. Secondly, we build the interaction relation estimation framework and multiperson relation graph network to represent the multiperson scene. Finally, we infered human behavior by the Petri network. In the following, Section 2 discusses overview of the proposed method. In Section 3, we discuss the multiperson interaction estimation framework, multiperson occlusion state reconstruction based on semantic knowledge event, and Petri Nets (PNs). Section 4 provides a comparative experiment. In the final section, general conclusions are made and possible further improvements on this research are stated.

2. Related Work

Multiperson interactive activity recognition is a comprehensive analytical task that has developed rapidly in recent years. In this section, we will briefly review the development of the multiperson interactive activity recognition. Recognizing the collective activity of a group of participators becomes an attractive research topic these years. Usually, multiperson interactive activity recognition needs to infer the complex interactions among different activity participators, and it is much more challenging than recognizing action of an individual subject.

2.1. Descriptor Learning without Interaction Modelling. At the early stage, researchers aim at seeking a discriminative descriptor to summarize the “multiperson interactive activity states” in a collective scenario. For this purpose, Zhao et al. [22] proposed a unified discriminative learning framework of multiple context models for concurrent collective activity recognition. Chen et al. [23] presented a new attribute-based spatio-temporal (AST) feature representation descriptor including spatio-temporal (ST) features and attribute features.

By means of the seminal works [24, 25], deep neural networks are used for extracting features with high representational capacity. Sudhakaran et al. [26] presented EgoACO for video action recognition that learns to pool action-context-object descriptors from frame level features by leveraging the verb-noun structure of action labels. By the aid of the multiscale feature maps output by a fully convolutional network, features of each individual within consecutive frames are fused together by a recurrent network [17].

2.2. Interaction Modelling with Shallow Models. Several works attempted to explore the interactions between individuals by shallow models. Dong et al. [27] proposed the Residual 3D Network (R3D) and Attention Residual 3D Network (AR3D) human action recognition models. Yang et al. [28] proposed a plug-and-play channel adaptive merging module (CAMM) based on graph convolutional networks (GCNs) specific for the human skeleton graph, which can merge the vertices from the same part of the skeleton graph adaptively and efficiently. Sheng et al. [29] proposed a discriminative subspace learning model (DLSM) to explore the complementary properties between the handcrafted shallow feature representations and the deep features. Song et al. [30] proposed a video-level 2D feature representation and a temporal attention model within a shallow convolutional neural network to efficiently exploit the temporal-spatial dynamics. Yan et al. [31] designed a Hierarchical Graph-based Cross Inference Network (HiGCIN), in which three levels of information include the body-region level, person level, and group-activity level. Li et al. [32] proposed symbiotic graph neural networks, which contain a backbone, an action-recognition head, and a motion-prediction head.

To sum up, these research studies intended to capture the interactive actions in collective activities by a shallow model, rendering them inapplicable for describing activities with complex interactions.

2.3. Deep Interaction Modelling. Zhou et al. [33] proposed the cascaded parsing network (CP-HOI) for a multistage, structured HOI understanding, which each cascade stage
refines HOI proposals and feeds them into a structured interaction reasoning module. Fan et al. [34] proposed a spatio-temporal graph neural network to represent the diverse gaze interactions and to infer atomic-level gaze communication by message passing. Qi et al. [35] proposed the graph parsing neural network (GPNN) to infer the HOI graph structure and the node labels. Rahmani et al. [36] proposed the robust nonlinear knowledge transfer model (R-NKTM) for human action recognition. Xiao et al. [37] proposed a dual attention network model which reasons about human-object interactions. This network weighs the important features for objects and actions, respectively.

Whereas the aforementioned method seems reasonable, it has several drawbacks. First of all, the majority of state-of-the-art detection methods do not use any kind of joint optimization to handle multiple objects, but rather rely on heuristic postprocessing. Thus, they are susceptible to greedy nonoptimal decisions. Secondly, the extracted features individually for each object discarded a large amount of context and interactions, which can be useful to reason the multiperson interactive behavior. This point is particularly important because of the locations and actions of some interactive humans which can be highly correlated. Thirdly, the independent detection and tracking methods mean that the representation used for localization is discarded, whereas re-using it would be more efficient. Finally, the sequential approach does not scale well with many people in the scene, as it requires multiples runs for a single image.

3. Overview of the Proposed Method

In order to overcome the above limitations and address specific problems, we propose the multiperson interactive relation graph (MRG) to simultaneously capture the interpersonal appearance and position relation. Firstly, we build the interaction relation estimation framework and multiperson relation graph network to represent the multiperson scene. Secondly, we discuss the multiperson interaction estimation framework. Section 4 discusses multiperson occlusion state reconstruction based on the semantic knowledge event. Finally, we introduce Petri Nets (PNs) for detecting activities in the TV human interaction dataset, Vlog, Atomic Visual Actions, and MLB-YouTube datasets. The system flowchart is shown in Figure 1.

3.1. Multiperson Interaction Estimation Framework. The framework is designed to recognize the multiperson interactive scene by clear-cut relation information and the input of extracted feature results and generate a set of reliable bounding box coordinates with their corresponding confidence scores via the target in the detection stage. On this basis, we build the interaction relation estimation framework and multiperson relation graph network to represent the multiperson scene. Responding to the self-occlusion issue in the field of human action recognition, a new adaptive occlusion state behavior recognition approach was presented based on the Petri network. In the following section, we will give detailed descriptions of our approach.

3.1.1. Interaction Estimation Framework. Firstly, in order to obtain multiperson objects from the surveillance video, we use the method [17] to target multiperson.

From being given the feature map \( F \in \mathbb{R}^{I \times D} \) and two dense maps \( B \in \mathbb{R}^{I \times 6} \) and \( P \in \mathbb{R}^{I \times P} \) (\( P \) represents a segmentation mask encoding in which parts of the image contain multiperson and \( B \) represents the coordinates of the bounding boxes of the people present in the scene, encoded relative to the pixel locations), we convert the given ground truth object locations into dense ground truth maps \( \widetilde{B} \) and \( \widetilde{P} \), for detecting the set of bounding boxes \( \{(x_{i0}, y_{i0}), (x_{i1}, y_{i1}), \ldots, (x_{in}, y_{in})\} \):

\[
\begin{align*}
    t_{y0} &= \frac{(i_{y} - y_{0})}{s_y}, \\
    t_{x0} &= \frac{(i_{x} - x_{0})}{s_x}, \\
    t_{y1} &= \frac{(y_{1} - i_{y})}{s_y}, \\
    t_{x1} &= \frac{(x_{1} - i_{x})}{s_x},
\end{align*}
\]

where \( s_y \) and \( s_x \) are scaling coefficients that are fixed as the maximum size of the bounding box over the training set image. \( \widetilde{B} \) are defined for \( i; \widetilde{P}_i = 1 \), and the regression loss is constructed accordingly. The loss is defined as follows:

\[
L_{det} = -\frac{1}{|T|} \sum_i \sum_{\tilde{B}_i} \log p_i + \frac{1}{\sum_{\tilde{P}_i} \sum_{\tilde{P}_i} ||\tilde{P}_i - \tilde{B}_i||_2^2},
\]

where \( w \) is a weight that makes training focused more on classification or regression for datasets where classification is easy, such as volleyball.

Secondly, we use the method [18] to acquire appearance relation and position relation:

\[
f_a(x^a_i, x^a_j) = \frac{(x^a_i)^T x^a_j}{\sqrt{d}},
\]

where \( \sqrt{d} \) acts as a normalization factor:

\[
f_s(x_i', x_j') = \prod(d(x_i', x_j') \leq \mu),
\]

where \( \prod(\cdot) \) is the indicator function; according to the actual scene, set a threshold value to determine 1 or 0. \( d(x_i', x_j') \) denotes the Euclidean distance between center points of multiperson bounding boxes, and \( \mu \) acts as a distance threshold which is a hyperparameter.

The parametric \( \xi \) is represented as the relative distance among multiperson position relation value using cosine and sine functions of different wavelengths. The feature dimension after embedding is \( d \). We then transform the embedded feature into a scalar by weight vectors \( W \) and \( b \), followed by a RELU activation:

\[
f_s(x_i', x_j') = \text{RELU}(W \xi(x_i', x_j') + b),
\]
In order to recognize the interaction relationship of multiperson behavior, we developed a multiperson relation method in which we defined three indexes [19]: movement time (MT), nonoverlapped movement time (NOMT), and group movement time (GMT). MT denotes the movement time of a single user, GMT is the total movement time of all persons, and NOMT is the nonoverlapped movement time between persons.

Generalizing the equation, in a given duration, if the MT of person is $MT_i$, the GMT of person $_1\cdot person_{n-1}$ is $GMT_1\cdot\cdots\cdot GMT_{n-1}$, and the NOMT of person $\cdot\cdot\cdot person_{n-1}$ is $NOMT_1\cdot\cdots\cdot NOMT_{n-1}$; then, GMT and NOMT of user 1, user 2, $\cdot\cdot\cdot$ user $n$ are calculated as below:

$$GMT_{1,2,\ldots,n} = GMT_{1,2,\ldots,n-1} + MT + (MT_n \cap GMT_{1,2,\ldots,n-1}),$$

$$NOMT_{1,2,\ldots,n} = NOMT_{1,2,\ldots,n-1} + MT_n + (MT_n \cap GMT_{1,2,\ldots,n-1}) + (MT_n \cap NOMT_{1,2,\ldots,n-1}).$$

(6)

Assuming that the higher the NOST, the more likely that there is an interaction, the networking method recognizes the interaction persons through four stages: (1) establish a potential interaction persons, (2) link another person to the potential interaction persons, (3) confirm an interaction relation, and (4) recognize another interaction relation. These stages are described in detail as follows:

Stage 1: establish the potential persons’ interaction relation

1. Calculate the NOMT of all linkable cases between any two persons with the highest NOMT.
2. If the two people’s NOMT is higher than respective MT, a potential interactive relation is identified. Otherwise, there would be none.

Stage 2: connect other people to the potential interactive relation

1. Calculate the NOMT of the potential one and any individual, and then, connect with a person who increases NOMT the most. After that, the potential multiperson interaction is re-established involving the newly connected person.
2. Repeat the above steps until there is no user left who can increase the NOMT. If there is any people who are not involved in the potential interaction relation, they are deemed to be noninteractive persons.

Stage 3: confirm the interactor

1. If GMT of the potential interaction relation is equal or more than a value, it is confirmed as an interaction
2. If GMT of the potential interaction relationship is less than a value, all of the collected relations are considered noninteraction

Stage 4: recognize group interactions

When they are more than two interactors, the networking methods can be utilized multiple times to recognize interactive relations.

As mentioned above, we utilize the graph structure to explicitly model pair-wise relation information for group activity understanding. Our design is inspired by the recent success of relational reasoning and graph neural networks [8, 38].

Formally, the nodes in our graph correspond to a set of actors $A = \{(GMT_i, NOMT_i) | i = 1, \ldots, n\}$, where $i$ is the number of actors, $x_i \in A$ is actor $i$’s appearance feature, and $x_i = (t_i, r_i)$ is the center coordinates of actor $i$’s bounding box. According to the existing method [38], we construct graph $G \in R^{N \times N}$ to represent pair-wise relation among actors, where relation value $G_{ij}$ indicates the importance of actor $j$’s feature to actor $i$: $G_{ij} = h(GMT_i(x_i^a, x_j^a), NOMT_i(x_i^a, x_j^a))$, (7)

where $GMT_i(x_i^a, x_j^a)$ denotes the appearance relation between two actors, and the position relation is computed by NOMT_i(x_i^a, x_j^a). The function $h$ fuses appearance and position relation to a scalar weight:

$$GMT_i(x_i^a, x_j^a) = \frac{(x_i^a)^T}{\sqrt{d}},$$

(8)

where $\sqrt{d}$ act as a normalization factor:

$$NOMT_i(x_i^a, x_j^a) = \prod(d(x_i^a, x_j^a) \leq \mu),$$

(9)

where $d(x_i^a, x_j^a)$ denotes the Euclidean distance between center points of two actors’ bounding boxes and $\mu$ acts as a distance threshold which is a hyperparameter.

3.1.2. Feature Description. We use a 3D extension to the SIFT algorithm [39, 40], as described in [41, 42], to determine the location of interest points. Given a 3D input volume $I(x, y, z)$ and a 3D Gaussian filter $G(x, y, z, \delta)$, we form multiscale difference of Gaussian (DoG) volumes, similar to [38, 39], as follows:

$$DoG(x, y, z, \delta) = I(x, y, z) \ast G(x, y, z, \delta^2),$$

$$- I(x, y, z) \ast G(x, y, z, \delta^{k-1}),$$

$$H = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{yx} & D_{yy} & D_{yz} \\ D_{zx} & D_{zy} & D_{zz} \end{bmatrix},$$

(10)

where $D_{ij}$ are the second derivatives in the DoG volume. Stavropoulos et al. [43] derived a measure to reject points using the trace and determinant of $H$, where

$$Tr(H) = D_{xx} + D_{yy} + D_{zz},$$

$$Det(H) = D_{xx}D_{yy}D_{zz} + 2D_{xy}D_{yz}D_{xz} - D_{xx}D_{yz}D_{zy} - D_{xx}D_{yz}D_{zy} - D_{xx}D_{yy}D_{xz} - D_{xx}D_{yy}D_{xz} - D_{xx}D_{yy}D_{xz} - D_{xx}D_{yy}D_{xz} - D_{xx}D_{yy}D_{xz} - D_{xx}D_{yy}D_{xz} - D_{xx}D_{yy}D_{xz},$$

(11)

If there are reject points, we define the following equation:
\[
\frac{\text{Tr}^3(H)}{\text{Det}(H)} > \left( \frac{2\tau_e + 1}{\tau_e} \right)^3.
\] (12)

Since the ultrasound image is noisy and not as sharp as the normal image, we lessen this threshold restriction so as to obtain more feature points [39]. \( \tau_e = 25 \), and this threshold is obtained from practical tests and has a certain universality. Therefore, we define the subvoxel estimate of the extrema true location is achieved by quadratic interpolation on the DoG volume data. After obtaining the identification of interest point locale, we define a localized neighborhood function, extending this from the earlier work [39, 40]:

\[
w(d, \delta) = \exp \left[ -\left( \frac{d}{\delta} \right)^2 \right],
\] (13)

where \( w(d, \delta) \) can limit the contribution of voxels around the point of interest to those in the local neighborhood, \( d \) is the voxel distance from the point of interest to the contributing voxel, and \( \delta \) is used to determine the extent of the local contribution:

\[
D_l = \frac{\sum_k w_k \cdot w(d_k, \delta)}{\sum_k w(d_k, \delta)}
\] (14)

For voxel \( k \), the voxel distance \( d_k \) from the adjacent acquired interest point location with the density \( \rho_k \).

### 3.2. Multiperson Occlusion State Reconstruction Based on Semantic Knowledge Event

This section presents the semantic knowledge techniques employed in the framework, namely, the semantic knowledge interpretation (SKI) components based on the method [25], which, respectively, provide a unanimous representation storage (representation layer) and semantic interpretation and event fusion (interpretation layer). The semantic-knowledge structures and vocabularies are described as the follows.

The semantic knowledge modelling and storage allows end users to model domain knowledge about (1) goal-oriented protocols, (2) domain observation entities and events, and (3) interactive occlusion behavior contextual models. Semantic knowledge of complex activities involved in each scenario.

The multiperson occlusion protocol (or scene) can be marked as an instance, which is applicable for the preservation of multiperson occlusion information of the occlusion state. The participant instances allow profile-related assertions about participants to be defined, such as interaction occlusion and self-occlusion. Protocol steps cover one or multiple tasks, with a start and an end node. Our proposed method implements three protocol steps: directional activities, occlusion activities, and self-occlusion activities. The term events suggest lower-level observation types and high-level activities. The ontology offers lightweight glossary for basic event-related information, for instance, event hierarchies and temporal extension. Event is the root class with two direct subclasses’ observation and activity for modelling observations and activities, respectively. The event has two embranchments: observations and activities. Four observation types derived from the observation modelling are, namely, modelling location (e.g., in the surveillance area), postures (e.g., standing up), actions (e.g., drinking water), and objects (e.g., a wallet). These are the basic types under the observation category.

The agents of the events and the temporal context are captured using constructs from DUL [43, 44] and OWL. Time [45], respectively. For example, the detection of the event object is modelled as follows:

```
: Purse1 a em: purse;
Leo@Time: t1;
Leo:involvedAgent [a du1: Agent].
:t1 a time: TemporalEntity
Time: hasBeginning [a time: Instant; time: inXSDDateTime]
Time: hasEnd
[A time: Instant;
Time: inXSDDateTime]
A time: Occlusion Moment;
Time: inXSDDateTime]
```

To perfect the definition of the interactive behavior structure in the occlusion state, so as to better share and reuse knowledge, this paper chooses an active ontology for the occluded interactive behavior modelling. The ontology is able to recognize the occlude activity and observation type, and the context of the occlusion and self-occlusion interactive activity thus could be stated. The knowledge representation module of this framework creates a model for semantic modelling in the context. Contextual information in each occlusion activity is conveyed by class equivalence axioms, which connects interactive behaviors with lower-level observations. More precisely, the activity models encompass the domain semantic knowledge required to detect complex activities. It is manifested by the mutual independence of lower-level observations and sophisticated behaviors and is composed of the following knowledge structure.

### 3.3. Action Recognition

In this paper, we introduce Petri Nets (PNs) for detecting activities in the TV human interaction dataset, Vlog, Atomic Visual Actions, and MLBYouTube datasets. The Petri Nets are graph-based techniques that can model and visualize various behavior types including parallelism, concurrency, resource sharing, and synchronization [46, 47]. A Petri Net is a finite state machine that allows multiple inputs and multiple outputs (a traditional finite state machine is a Petri Net in which each transition is restricted to have exactly one output and one input [48]. In a graphical representation [49–51], the places are drawn as circles and the transitions are drawn as squares or rectangles. Arcs are drawn connecting place nodes to transition nodes (input arcs) or transition nodes to place nodes (output arcs). Regular arcs are drawn with arrow heads. Inhibit arcs are drawn with dot heads. Arcs are
associated with a weight, also called the arc’s multiplicity. Arc multiplicity is taken to be one if not specified. Places connected to a transition by input arcs are called transition’s input places (or input set). Similarly, places connected to a transition by output arcs are called transition’s output places (or output set). A place node may contain a number of tokens (another graph component). Tokens are visualized as black dots within the place node which contains them.

As PN is introduced to recognize the relative velocity of area models and nodes, the proposed method is based on a primitive advanced Petri Net, also known as the PN with area-velocity tokens.

Firstly, define a Dong et al. [27]-like model and reasoning as follows.

With regard to the format, the basic place/transition PN can be described as a five-tuple: PN = \( \{ P, T, I, O, M \} \) and can be graphically represented by a directed bipartite graph, which includes two types of nodes: the places \( P \), which are drawn as circles, and the transitions \( T \), which are drawn either as bars or boxes [15]:

\[
P = \{ p_1, p_2, p_3, \ldots, p_m \} \text{ is a finite set of places}
\]

\[
T = \{ t_1, t_2, t_3, \ldots, t_m \} \text{ is a finite set of transitions}
\]

\[
I: \ (P \times T) \rightarrow N \text{ is the input arc function which can be represented by the input matrix: } I_{\text{nom}}. \text{ If there existed an arc with weight } k \text{ that connects the place } p_i \text{ to the transition } t_j, \text{ then } I(p_i, t_j) = k; \text{ otherwise, } I(p_i, t_j) = 0.
\]

\[
O: \ (P \times T) \rightarrow N \text{ is the output arc function, which can be represented by the output matrix: } O_{\text{nom}}. \text{ If there exists an arc with weight } w \text{ that connects the transition } t_j \text{ to the place } p_k, \text{ then } O(t_j, p_k) = w; \text{ otherwise, } O(t_j, p_k) = 0.
\]

\[
M: \ P \rightarrow N \text{ is the current marking of the net and can be represented as a vector } M_{\text{1x}m}. \text{ } M_0 \text{ is the initial marking, which denotes the initial state of the net.}
\]

After a new transformed function, the behavior recognition is as follows:

\[
X_{\text{new}}^i = X_{\text{old}}^i + S_j \cdot T_j,
\]

where \( S_j \) is the detector response for the action \( j \), \( T_j \) is the area-velocity goodness of the transition \( j \), and \( X_{\text{old}}^i \) is the previous score of the token \( i \) that is passing through the transition \( j \).

After firing a given transition (rule), tokens from its input places are removed. Firing a given transition (rule) and removing tokens can be intuitively interpreted as an execution reasoning by using this rule in a given reasoning process. Hence, in the next steps, markings of input places of a fired rule are already unnecessary. Such reasoning can be understood as a kind of forward reasoning.

4. Experiments

Results of the TV human interaction dataset [52] are shown in Figure 2. The Interaction dataset consists of 300 video clips collected from over 20 different TV shows, containing four types of interactions: hand shakes, high fives, hugs, and kisses. Noninteraction clips are also attached.

Vlog [53] is a large-scale YouTube dataset, aiming to collect lifestyle vlogs which includes daily interactions between people and things. The experimental results of this dataset are shown in Figure 3.

Atomic Visual Actions [54] is a dataset released by UC Berkeley and Google with a diverse environment and a large number of classes labelled using exhaustive frame-level annotation of YouTube videos. 430 clips were extracted from the 15 to 30 mins from movies and television shows related to famous actors from different countries. The results of the Atomic Visual Actions dataset are shown in Figure 4.

The MLB-YouTube (Major League Baseball) [55] is a fine-grained activity dataset consisting of 20 baseball games from the 2017 MLB postseason. Among them, the 9 activities (swing, foul, ball, and strike and no action) are not very distinct whereas salient repetition of activities are noticed. The annotations are multilabel and overlapping along with pitch type (e.g., fastball, curveball, and slider) and the speed of the pitch also being given for each pitch. The results of the MLB-YouTube (Major League Baseball) are shown in Figure 5. The experimental results are shown in Tables 1–4, respectively.

It can be seen from the above table that the average recognition accuracy of the recognition method in this section is 93.42%, which is superior to other recognition algorithms.

It can be seen from the above table that the average recognition accuracy of the recognition method in this section is 93.37%, which is superior to other recognition algorithms.
It can be seen from the above table that the average recognition accuracy of the recognition method in this section is 93.21%, which is superior to other recognition algorithms.

It can be seen from the above table that the average recognition accuracy of the recognition method in this section is 92.87%, which is superior to other recognition algorithms.

Hence, according to the above tables, our approach proved to be more accurate in recognizing human-computer interactions, which further validates the accuracy and efficiency of our algorithm.

5. Conclusions and Future Work

The present study proposes a newly-developed interactive behavior relational modelling and identification framework, which successfully recognizes multiperson interaction
behaviors. We build a flexible and efficient model of the multiperson interaction relation estimation framework to capture the appearance and position relation of people, and a new reasoning approach is thus formed. The temporal consistency is handled via a person-level matching recurrent neural network. With regard to the integration of the interaction relation estimation framework and multiperson relation graph network, it not only facilitates automatic learning of the human interaction dataset in an end-to-end manner in our proposed approach but also enables that the reasoning process could be efficiently performed with standard matrix operations. As for the self-occlusion issue in human action recognition, this paper proposes a new adaptive occlusion state behavior recognition approach based on semantic knowledge event representation.

On account of four standard interactive behavior database, we visualize the learned interaction relation estimation framework and the multiperson relation graph (MRG), which demonstrates that the proposed method is able to capture the discriminative relation information for multiperson interactive activity recognition. Experimental results reveal that the multiperson occlusion state reconstruction based on SKE outperforms other approaches in its accuracy.

There are some failure cases in all of the above experiment. During the experiment process, if there are some strong occlusion (e.g., the occlusion area is over one-third) and high dynamic videos (e.g., interactive actions switch quickly. For instance, the cricket is quickly dropped from the hand) among multiperson or person objects, the identification accuracy would significantly decrease. Because of strong occlusion and high dynamic, it can cause feature points and interaction relationship positioning to be inaccurate and location can only be expected.

We extracted the above database containing video data with strong occlusion and high dynamic video and conducted experiments with the proposed method in the paper. The results are shown in Tables 5–8.

It can be seen from the above table that the average recognition accuracy of the recognition method in this section is 30.26%, which is superior to other recognition algorithms.

It can be seen from the above table that the average recognition accuracy of the recognition method in this section is 31.12%, which is superior to other recognition algorithms.

It can be seen from the above table that the average recognition accuracy of the recognition method in this section is 34.23%, which is superior to other recognition algorithms.

It can be seen from the above table that the average recognition accuracy of the recognition method in this section is 28.12%, which is inferior to other recognition algorithms.

It is assumed that the proposed method can be extended to interacting people or objects. Tracking poses of interacting people, however, will involve more complex factors for instance, dealing with more variable motion, interperson occlusions, and possible appearance similarity of different people. [51]

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported in part by the Project of Science and Technology Department of Jiangxi Province under Grant 20181BAB202017 and 20202BAB212004, Huxiang high-level talent gathering project 2019RS1072, and Science and Technology Research Project of Jiangxi Provincial Educational Department under Grant GJJ170373 and GJJ190354.

References

[1] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, “A hierarchical deep temporal model for group activity recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, June 2016.

[2] P. Zhang, Y. Tang, J.-F. Hu, and W.-S. Zheng, “Fast collective activity recognition under weak supervision,” IEEE Transactions on Image Processing, vol. 29, no. 1, pp. 29–43, 2020.

[3] V. Ramanathan, J. Huang, S. Abu-El-Haija, A. Gorbans, K. Murphy, and L. Fei-Fei, “Detecting events and key actors in multi-person videos,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, June 2016.

[4] R. Girshick, “Fast r-cnn,” in Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448, Santiago, Chile, December 2015.

[5] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: towards real-time object detection with region proposal networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, 2017.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: unified, real-time object detection,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, June 2016.

[7] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, June 2016.

[8] Y. Peng, H. Zhong, Z. Xu, H. Tu, X. Li, and L. Peng, “Pulmonary lobe segmentation in CT images based on lung anatomy knowledge,” Mathematical Problems in Engineering, vol. 2021, Article ID 5588629, 15 pages, 2021.

[9] X. Xiaobin Chang, W. S. Wei-Shi Zheng, and J. Jianguo Zhang, “Learning person-person interaction in collective activity recognition,” IEEE Transactions on Image Processing, vol. 24, no. 6, pp. 1905–1918, 2015.

[10] H. Wu, S. Jie, X. Xing, Y. Ji, and H. T. Shen, “Recognition and detection of two-person interactive actions using automatically selected skeleton features,” IEEE Transactions on Human-Machine Systems, vol. 48, pp. 1–7, 2017.
[11] L. An-An, X. Ning, and W. Nie, “Multi-domain and multi-task learning for human action recognition,” IEEE Transactions on Image Processing, vol. 28, pp. 853–867, 2018.

[12] Z. Zhu, G. D. Guglielmo, Q. Cheng et al., “Photonic switched optically connected memory: an approach to address memory challenges in deep learning,” Journal of Lightwave Technology, vol. 38, no. 10, pp. 2815–2825, 2020.

[13] D. Xu, H. Fu, L. Wu, M. Jian, D. Wang, and X. Liu, “Group activity recognition by using effective multiple modality relation representation with temporal-spatial attention,” IEEE Access, vol. 8, pp. 65689–65698, 2020.

[14] Y. Kim, A. Ghorpade, F. Zhao, F. C. Pereira, P. C. Zegras, and M. Ben-Akiva, “Activity recognition for a smartphone and web-based human mobility sensing system,” IEEE Intelligent Systems, vol. 33, no. 4, pp. 5–23, 2018.

[15] J. Liu, C. Wang, Y. Gong, and H. Xue, “Deep fully connected model for collective activity recognition,” IEEE Access, vol. 7, pp. 104308–104314, 2019.

[16] L. Lu, R. Yu, H. Di, L. Zhang, and Y. Lu, “GAIM: graph attention interaction model for collective activity recognition,” IEEE Transactions on Multimedia, vol. 22, no. 2, pp. 524–539, 2019.

[17] T. Bagautdinov, A. Alahi, F. Fleuret, P. Fua, and S. Savarese, “Social scene understanding: end-to-end multi-person action localization and collective activity recognition,” in Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3425–3434, Honolulu, HI, USA, July 2017.

[18] J. Wu, L. Wang, L. Wang, J. Guo, and G. Wu, “Learning actor relation graphs for group activity recognition,” 2019, https://arxiv.org/abs/1904.10117.

[19] J. H. Ma and S. H. Cha, “A human data-driven interaction estimation using IoT sensors for workplace design,” Automation in Construction, vol. 119, pp. 1–12, 2020.

[20] Z. Hossein-Nejad, H. Agahi, and A. Mahmoodzadeh, “Image matching based on the adaptive redundant key point elimination method in the SIFT algorithm,” Pattern Analysis & Applications, vol. 24, pp. 1–15, 2020.

[21] D. Lowe, “Distinctive image features from scale-invariant keypoints,” International Journal of Computer Vision, vol. 60, pp. 91–110, 2014.

[22] C. Y. Zhao, J. Q. Wang, and H. Q. Lu, “Discriminative context models for collective activity recognition,” In Multimedia Tools and Applications, vol. 76, no. 5, pp. 648–653, 2014.

[23] C. Chen, H. Dou, and Z. Gan, “Collective activity recognition by attribute-based spatio-temporal descriptor,” IEICE Transactions on Info and Systems, vol. E98.D, no. 10, pp. 1875–1878, 2015.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Advances in Neural Information Processing Systems, vol. 26, pp. 1097–1105, 2012.

[25] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778, Las Vegas, NV, USA, June 2016.

[26] S. Sudhakaran, S. Escalera, and O. Lanz, “Learning to recognize actions on objects in egocentric video with attention dictionaries,” 2021, https://arxiv.org/abs/2102.08065.

[27] M. Dong, Z. Fang, Y. Li, S. Bi, and J. Chen, “AR3D: attention residual 3D network for human action recognition,” Sensors, vol. 21, no. 5, p. 1656, 2021.

[28] W. Yang, J. Zhang, J. Cai, and Z. Xu, “Shallow graph convolutional network for skeleton-based action recognition,” Sensors, vol. 21, no. 2, p. 452, 2021.

[29] B. Sheng, J. Li, X. Fu, Q. Li, and J. Han, “Discriminative multi-view subpace feature learning for action recognition,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 99, p. 1, 2019.

[30] X. Song, C. Lan, W. Zeng, J. Xing, X. Sun, and J. Yang, “Temporal-spatial mapping for action recognition,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 3, pp. 748–759, 2020.

[31] R. Yan, L. Xie, J. Tang, X. Shu, and Q. Tian, “HiGCIN: hierarchical graph-based cross inference network for group activity recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–14, 2020.

[32] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, “Symbiotic graph neural networks for 3D skeleton-based human action recognition and motion prediction,” IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–19, 2021.

[33] T. Zhou, S. Qi, W. Wang, J. Shen, and S.-C. Zhu, “Cascaded parsing of human-object interaction recognition,” IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 1–14, 2021.

[34] L. Fan, W. Wang, S.-C. Zhu, X. Tang, and S. Huang, “Understanding human gaze communication by spatio-temporal graph reasoning,” in Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, November 2019.

[35] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, Learning Human-Object Interactions by Graph Parsing Neural Networks, Springer Science+Business Media, Berlin, Germany, 2018.

[36] H. Rahmani, A. Mian, and M. Shah, “Learning a deep model for human action recognition from novel viewpoints,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 3, pp. 667–681, 2018.

[37] T. Xiao, Q. Fan, D. Gutfrudend, M. Monfort, A. Oliva, and B. Zhou, “Reasoning about human-object interactions through dual attention networks,” 2019 IEEE/CVF International Conference on Computer Vision (ICCV), in Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3918–3927, Seoul, Republic of Korea, November 2019.

[38] J. Wu, L. Wang, L. Wang, J. Guo, and G. Wu, “Learning actor relation graphs for group Activity recognition,” in Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9956–9966, Long Beach, CA, USA, June 2019.

[39] D. Ni, Y. P. Chui, X. Qu et al., “Reconstruction of volumetric ultrasound panorama based on improved 3D SIFT,” Computerized Medical Imaging and Graphics, vol. 33, no. 7, pp. 559–566, 2009.

[40] S. Allaire, J. J. Kim, S. L. Breen, D. A. Jaffray, and V. Pekar, “Full orientation invariance and improved feature selectivity of 3D SIFT with application to medical image analysis,” in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–8, Anchorage, AK, USA, June 2008.

[41] M. Mesmakhosroshahi and J. Kim, “Improving spatio-temporal feature extraction techniques and their applications in action classification,” in Proceedings of the Visual Communications and Image Processing (VCIP), pp. 1–6, San Diego, CA, USA, November 2012.
[42] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional SIFT descriptor and its application to action recognition," in Proceedings of the 15th international conference on Multimedia, pp. 357–360, Augsburg, Germany, September 2007.

[43] T. G. Stavropoulos, G. Meditskos, S. Andreidis et al., "Semantic event fusion of computer vision and ambient sensor data for activity recognition to support dementia care," Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 8, pp. 3057–3072, 2020.

[44] S. Borgo and E. Blanzieri, "Trait-based module for culturally-competent robots," In International Journal of Humanoid Robotics, vol. 16, no. 06, pp. 358–365, 2013.

[45] M. A. Marsan, G. Balbo, G. Conte, S. Donatelli, and G. Franceschinis, "Modelling with generalized stochastic petri nets," ACM SIGMETRICS-Performance Evaluation Review, vol. 26, no. 2, p. 2, 1998.

[46] A. Patron-Perez, TV Human Interaction Dataset, https://www.robots.ox.ac.uk/%7Ealonso/tv_human_inter%20actions.html, 2010.

[47] D. F. Fouhey, W.-C. Kuo, A. A. Efros, and J. Malik, "From lifestyle Vlogs to everyday interactions," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 4991–5000, Salt Lake City, UT, USA, June 2018.

[48] C. Gu, C. Sun, D. A. Ross et al., "AVA: a video dataset of spatio-temporally localized atomic visual actions," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 6047–6056, Salt Lake City, UT, USA, June 2018.

[49] A. J. Piergiovanni and M. S. Ryoo, "Fine-grained activity recognition in baseball videos," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 6047–6056, June 2018.

[50] M. Afsariabi, H. Khotanlou, and M. Mansoorizadeh, "DTW-CNN: time series-based human interaction prediction in videos using CNN-extracted features," The Visual Computer, vol. 36, no. 6, pp. 1127–1139, 2019.

[51] A. M. Atto, A. Benoit, and P. Lambert, "Timed-image based deep learning for action recognition in video sequences," Pattern Recognition, vol. 104, pp. 1–13, 2020.

[52] D.-H. Yoon, N.-G. Cho, and S.-W. Lee, "A novel online activity detection framework from untrimmed video streams," Pattern Recognition, vol. 106, pp. 107396–107411, 2020.

[53] Z. E. Wemlinger and L. B. Holder, "Cross-environment activity recognition using a shared semantic vocabulary," Pervasive and Mobile Computing, vol. 51, no. 51, pp. 150–159, 2018.

[54] T. Bagautdinov, A. Alahi, F. Fleuret, P. Fua, and S. Savarese, “Social scene understanding: end-to-end multi-person action localization and collective activity recognition," in Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3425–3434, Honolulu, HI, USA, July 2017.

[55] T. Hussain, K. Muhammad, A. Ullah, Z. Cao, S. W. Baik, and V. H. C. De Albuquerque, “Cloud-assisted multiview video summarization using CNN and bidirectional LSTM,” IEEE Transactions on Industrial Informatics, vol. 16, no. 1, pp. 77–86, 2020.

[56] S. U. Khan, I. U. Haq, S. Rho, S. W. Baik, and M. Y. Lee, “Cover the violence: a novel deep-learning-based approach towards violence-detection in movies,” Applied Sciences, vol. 9, no. 22, p. 4963, Article ID 4963, 2019.