Multiple Granularity Group Interaction Prediction

Taiping Yao*, Minsi Wang* , Bingbing Ni*†, Huawei Wei, Xiaokang Yang
Shanghai Institute for Advanced Communication and Data Science,
Shanghai Key Laboratory of Digital Media Processing and Transmission,
Shanghai Jiao Tong University, Shanghai 200240, China
sndlytp@sjtu.edu.cn, mswang1994@gmail.com
nibingbing@sjtu.edu.cn, weihuawei26@gmail.com, xkyang@sjtu.edu.cn

Abstract

*Most human activity analysis works (*i.e.*, recognition or prediction) only focus on a single granularity,* i.e.*, either modelling global motion based on the coarse level movement such as human trajectories or forecasting future detailed action based on body parts' movement such as skeleton motion. In contrast, in this work, we propose a multi-granularity interaction prediction network which integrates both global motion and detailed local action. Built on a bi-directional LSTM network, the proposed method possesses between granularities links which encourage feature sharing as well as cross-feature consistency between both global and local granularity (e.g., trajectory or local action), and in turn predict long-term global location and local dynamics of each individual. We validate our method on several public datasets with promising performance.*

1. Introduction

In collective activities, predicting multi-person interaction in multi-granularity including action and trajectory details of each individual is a challenging problem. It has many applications, such as group activity analysis, social event prediction and collective activity recognition.

Most works for human motion prediction mainly focus on a single person in a single granularity. Namely, they either forecast human activity by analyzing the action information of each person, or predict human movement only focusing on its trajectory. Previous works [19, 4, 15] utilize deep RNNs to model human dynamics. These works only consider local human action, but without global motion, *i.e.*, trajectory. Other works [2, 25, 29] only focus on human trajectory prediction in crowd spaces without considering the information of detailed action. All above works utilize only one type of information to independently analyze the complex human activity. However, global and local information (*i.e.*, trajectory and human action) describe the human activity from different views. Only using one of these information can not comprehensively represent the activity, especially in details. We believe that considering both granularities, *i.e.*, global and local information, and their interactions, can definitely help action analysis, i.e., prediction.

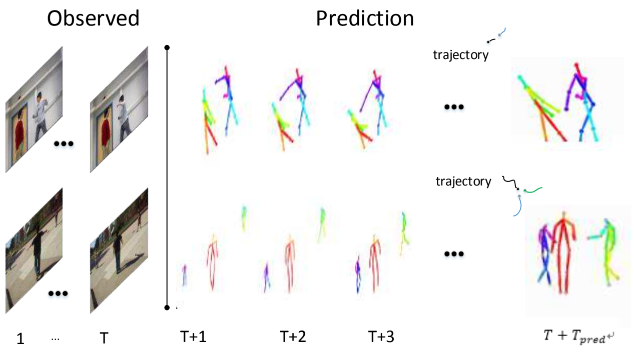

Figure 1. Given a short video clip, we propose a method to predict future group interaction in multi-granularity, including global motion (*trajectory*) and local motion (*part movement*)

We present the first study on multi-person interaction prediction focusing on multi-granularity. To predict the action and trajectory details of each individual involved in a group activity, we propose a method considering both global and local granularities, *i.e.*, trajectory and detailed body part movements. As shown in Figure.1, taking a short video clip with various number of persons as input, we aim to predict a sequence of future skeletal motion data and trajectory for all individuals.

This multi-person interaction prediction in a panoramic view is a challenging task. On the one hand, existing works only focus on a single granularity, but the information in single granularity is not sufficient to represent individual

*Equal Contribution.
†Corresponding Author.

dynamics in group activity. Namely, predicting group interaction in multi-granularity must take information of different views into consideration, including coarse level movement and fine level actions. Since information between different granularities is related, it is significant to model their interactions. On the other hand, in group activities, the action of individuals and positional information of the next frame are interacted with both their own dynamic information and the information of people associated. As collective activities always contain a varying number of people, the proposed network must be flexible enough to model mutual interactions simultaneously.

To address these issues, we propose a multi-granularity group interaction prediction architecture, which features a single-granularity prediction LSTM (sLSTM) combined with a novel multi-granularity interaction network. This network contains **intra**-granularity interaction sub-network and **inter**-granularity interaction sub-network. *i.e.*, the proposed network focuses on different granularities including trajectory and detailed action. Each sLSTM contains an encoder to capture spatio-temporal continuity. Then to model the interaction between different individuals within granularity, we propose two intra-granularity interaction sub-networks to model interaction in trajectory and action respectively. Further more, to model the interaction between different granularities, an inter-granularity interaction sub-network based on bi-directional LSTM is employed for its capability of preserving long memory in two directions. The proposed method has been comprehensively evaluated on several public datasets with three evaluation metrics. Experimental results demonstrate that our method can well address group interaction prediction problem.

2. Related work

Group Activity Analysis. Previous works on group activity analysis usually focus on group activity recognition. Lan *et al.* [17] proposed an adaptive latent structure learning recognizing group activities which jointly captures group activity, individual actions, and interactions among them. Social roles in [21] and [16] were proposed as the expected behavior of an individual in the context of a group. Choi and Savarese [9] unified tracking multiple people, recognizing individual actions, interactions and collective activities in a joint framework. In the work of Ibrahim *et al.* [13], a hierarchical deep temporal model has been used to aggregate person-level information for whole activity understanding. Bagautdinov *et al.* [3] have unified locations of individuals, social actions and collective activities in an end-to-end framework. Wang *et al.* [28] unified the interactional feature modeling process for single person dynamics, intra-group and inter-group interactions utilizing LSTM. Shu *et al.* [23] proposed a confidence-energy recurrent network to recognizing human activities at distinct semantic levels. All of above methods only focus on coarse-grained recognition, including recognizing individual actions, interactions, and collective activities. However, it is not sufficient to understand group activity and perform reasonable activity prediction only based on recognition results. To address this issue, we propose to predict multi-granularity group interaction which is more challenging than recognition.

Multi-granularity Analysis. Multi-granularity analysis has been successfully applied in many tasks including tracking, segmentation and classification [26, 31, 8, 32]. Wang *et al.* [26] proposed a fine-grained categorization framework trained from multiple granularity labels, and the results outperform most of the existing approaches. Yang *et al.* [31] introduced a Multiple Granularity Analysis framework for video segmentation in a coarse-to-fine manner. Chen *et al.* [8] proposed to use multi-granularity topics to generate features for short text, and this method can significantly reduce the classification errors. Multi-granularity embedding method proposed in [32] has been proved to improve word embedding by further leveraging both characters and radicals. The success of multi-granularity method inspires us to model group interaction in the multi-granularity manner for human motion prediction.

Human Motion Prediction. Modelling human motion plays an important role in many tasks including activity recognition [22, 34], motion generation [19, 15] and robotics [5]. Prior works have addressed this problem by using Hiden Markov Models(HMMS) [18], Gaussian process [27], restricted Boltzman machine(CRBM)[24] and random forest [10]. Recently, deep recurrent neural networks (RNN) have shown its superiority in sequence learning. Fragkiadaki *et al.* [11] proposed an ERD network. Jain *et al.* [15] proposed a method combining spatiotemporal graphs with RNNs. Martinez *et al.* [19] introduced a sequence-to-sequence model using a residual architecture which has obtained state-of-the-art performance. And Butepage *et al.* [4] used deep learning frameworks to extract deep feature representation for human motion prediction. Most of the above methods use joint angle data from H3.6M [14], which limits the development of human analysis. To deal with this problem, we propose a multi-granularity data generator to directly process on the 2D coordinates. Besides, although these previous works show satisfactory performance in modelling single person dynamics, the multi-person prediction problem involving interactions has not been well addressed. To the best of our knowledge, we are the first study focus on multi-granularity group interaction prediction.

3. Methodology

Previous human activity analysis works only consider single granularity, either focus on trajectory or focus on action. However single granularity information can not com-

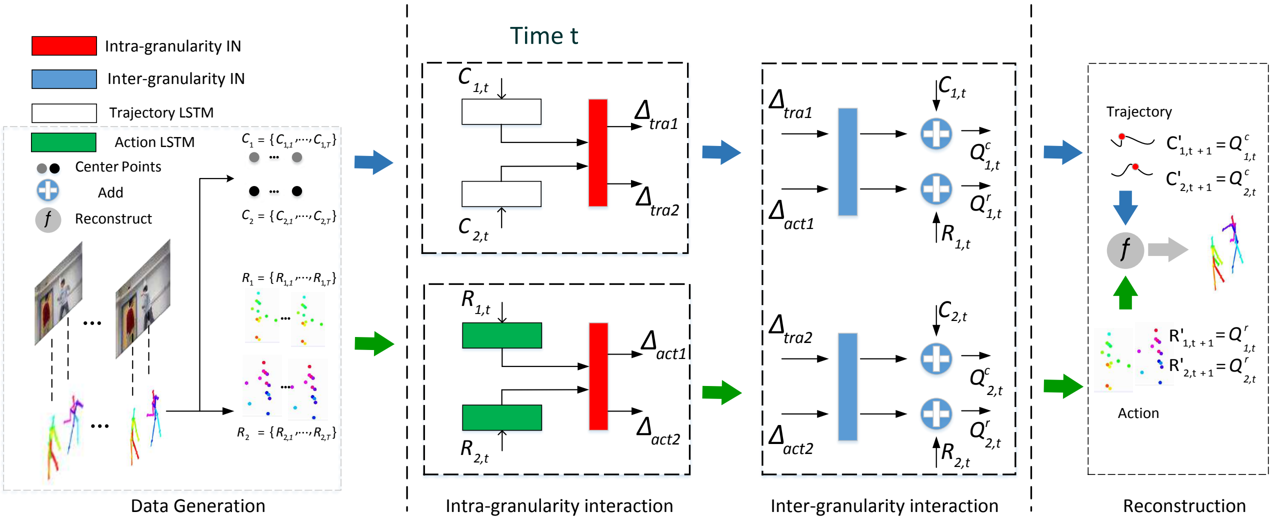

Figure 2. Overview of our multi-granularity interaction network at t time step, Δ_{act} is the change value of action and Δ_{tra} is the change value of trajectory. We use intra-granularity interaction sub-network and inter-granularity interaction sub-network to model interaction and predict trajectory and local body part's action. After combining them, our method outputs the generated skeleton data for next time step.

pletely represent the activity information. Additionally, different from single-person activity, individuals are interrelated and play different social roles in group activity. This motivates us to propose a multi-granularity interaction network to focus on multiple granularity and model the interaction between them.

As illustrated in Figure.2, proposed network consists of three parts. First, an input data generator splits every individual skeleton into a center point representing global location and other relative points representing local action details. Then, an intra-granularity interaction sub-network and an inter-granularity interaction sub-network are deployed to capture the interaction within granularity and between different granularity respectively, and predict different granularity information including trajectory and action for the next frame. In the end, we combine the trajectory with action information to obtain final actual skeletons.

3.1. Multi-granularity Data Generator

In group activity, each individual is involved with two basic granularities. One is the location in frame representing the spatial information. And the other is local action represented by skeletons. Previous works which model human motion or interaction, only pay attention to single granularity. However, both granularities are essential for correctly understanding the activity and precisely predicting future motion.Thus we propose a multi-granularity data generator to gain different granularity features.

The multi-granularity data generator takes a short video clip as input, crops individuals according to bounding box, then uses pose estimation method [6] to extract coordinates of skeleton key joints, and splits every individual skeleton into a center point and other relative points. First, utilizing human pose estimation network, each individual is represented by a feature vector which consists of xy-coordinates of K skeleton key points in every frame. Then, supposing that $S_{n,t} = \{X^k_{n,t}\}, k = 1, 2, ..., K$ denotes the skeleton data of n-th individual at t time step, and $X^k_{n,t}$ denotes the xy-coordinates of the k-th key point. For modelling multi-granularity, we select the center point in individual skeletons to represent the spatial information, which is usually the center of body. And other key points in skeleton are represented by the coordinates relative to the center point, which contain the information of individual action. Suppose we choose $X^c_{n,t}$ as the center point, and use following equations to represent the position $C_{n,t}$ and action $R_{n,t}$ respectively:

$$C_{n,t} = X^c_{n,t}, \tag{1}$$

$$R_{n,t} = \{X^k_{n,t} - C_{n,t}\}, k = 1, 2, ...K, k \neq c, \tag{2}$$

where $C_{n,t}$ is the absolute coordinates of center point and $R_{n,t}$ denotes the rest coordinates relative to $C_{n,t}$ at t time step for n-th individual. We standardize $R_{n,t}$ by mean subtraction and division by the standard deviation along each dimension. And $C_{n,t}$ subtracts the mean of all individual center coordinates to preserve relative spatial location information. After obtaining $C_{n,t}$ and $R_{n,t}$, the features in two granularities, we apply proposed multi-granularity interaction network for group activity prediction.

3.2. Multi-granularity Interaction Prediction Network

As mentioned above, individual features are partitioned into two granularities. Because individuals are interrelated in group activities, it is then important to model the interaction in granularity features, including intra-granularity interaction (i.e., trajectory to trajectory, action to action) and inter-granularity interaction (i.e., trajectory to action). To address this issue, we propose two different interaction subnetworks to capture these interaction, while also predict the features (i.e., trajectory and action) for the next time step.

3.2.1 Granularity Feature Encoding

As illustrated in Figure.2, two single-granularity LSTMs (**sLSTM**) are proposed to encode different granularity features for each individual. One encodes the trajectory features and the other encodes the action features. We denote by $C_{n,t}$ and $R_{n,t}$ the trajectory and action features of n-th individual at t time step respectively. Then in encode-stage, each single-granularity LSTM encodes the input ($C_{n,t}$ or $R_{n,t}$), and returns the output ($o^c_{n,t}$ or $o^r_{n,t}$) and hidden state ($h^c_{n,t}$ or $h^r_{n,t}$) at t time step for n-th individual, as Equation follows:

$$[o^c_{n,t}, h^c_{n,t}] = \psi_1(C_{n,t}, h^c_{n,t-1}), \tag{3}$$

$$[o^r_{n,t}, h^r_{n,t}] = \psi_2(R_{n,t}, h^r_{n,t-1}), \tag{4}$$

where ψ_1, ψ_2 denote trajectory sLSTM unit and action sLSTM unit. $o^c_{n,t}$ and $o^r_{n,t}$ denote the corresponding outputs of the two sLSTMs. We suppose that there are N individuals in the group activity at t time step, $O^c_t = \{o^c_{n,t}\}, n = 1, 2, ..., N$ denotes the encoded trajectory features of all individuals at t time step, Similarly, $O^r_t = \{o^r_{n,t}\}$ denotes the encoded action features of all individuals at t time step. Then O^c_t and O^r_t are taken as the input of the inra-granularity interaction network to capture the interaction between features in the same granularity.

3.2.2 Intra-granularity Interaction Network

In group activity, individuals adjust their paths and actions by reasoning about other individuals. Thus, after obtaining encoded features in different granularities, we need to capture the interaction of individuals in various granularity. First, we propose intra-granularity interaction network to capture the implicit interaction between features in the same granularity. More specifically, we use two intra-granularity interaction networks including **trajectory interaction subnetwork** and **action interaction sub-network** to capture the interaction between trajectories and the interaction between actions respectively.

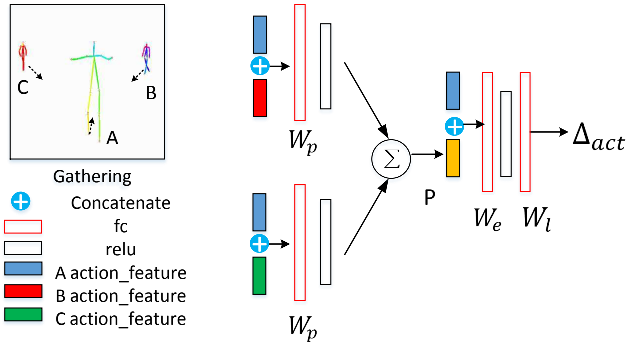

Figure 3. Intra-granularity interaction network

We propose the intra-granularity interaction sub-network inspired by [7], and the two intra-granularity interaction sub-networks have the same structure except for different dimensions of parameters. As illustrated in Figure.3, the intra-granularity interaction sub-network is composed of an encoder and a decoder. Supposing there are N individuals in a group activity and $o^c_{n,t}$, $o^r_{n,t}$ denote the encoded trajectory features and encoded action features of n-th individual at t time step. When modelling the impact from other individuals to n-th individual in action granularity at t time step, the input to the interaction sub-network is a vector consist of $N-1$ pairwise concatenation, $\{[o^r_{1,t}, o^r_{n,t}], [o^r_{2,t}, o^r_{n,t}], ...\}$. Most previous works only model the interaction between adjacent individuals instead of all individuals [2, 7]. However, our target is more focus on interactive pattern learning, thus we take all individuals into consideration. After obtaining the input, the encoder encodes all the pairwise concatenation and sums all the encodings as the whole impact P from other individuals to the n-th individual. Then whole impact P is concatenated with $o^r_{n,t}$ as the input to the decoder, and the decoder predicts the action change value Δ_{act} of the n-th individual for $t+1$ time step. The whole process is as follows:

$$p_i = \phi_1(o^r_{i,t}, o^r_{n,t}; W_p), i \neq n, \tag{5}$$

$$P = \sum_{i=1, i\neq n}^{N} p_i, \tag{6}$$

$$\Delta_{act} = \phi_2(\phi_1(P, o^r_{n,t}; W_e); W_l) \tag{7}$$

where $\phi_1(.)$ is an embedding function with RELU nonlinearity, $\phi_2(.)$ is an common embedding function. W_p,W_e and W_l are embedding weights. At last, the action interaction network predicts the action change value Δ_{act} for all individuals. Similarly, the trajectory interaction network predicts the trajectory change value Δ_{tra} for all individuals.

3.2.3 Inter-granularity Interaction Network

Intra-granularity interaction sub-network only captures the interaction between features in the same granularity without considering the relationship between features in different granularities, *e.g*. interaction between trajectory and action. To address this issue, we propose an inter-granularity interaction sub-network to model the cross-granularity interaction. Instead of directly modelling the interaction between trajectory features and action features, we model the interaction between Δ_{act} and Δ_{tra}. On the one hand, it helps avoid memorizing the environment. On the other hand, the relationship between original trajectory features and action features is hard to model. However, the change of trajectory can represent the general motion trend of the action. Correspondingly, the change of action can reflect the motion of trajectory to some extent. Thus modelling the relationship between Δ_{act} and Δ_{tra} is more reasonable.

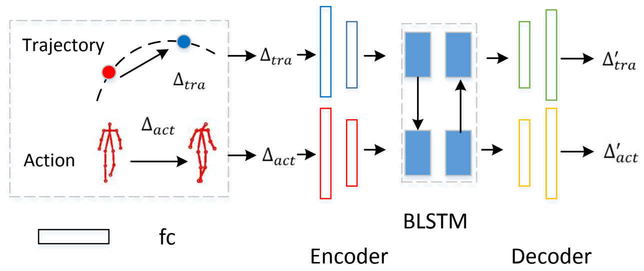

Figure 4. Inter-granularity interaction network

The inter-granularity interaction sub-network takes the prediction of intra-granularity interaction sub-network as input. For an individual, firstly, Δ_{act} and Δ_{tra} are embedded into a same latent space using different encoders. Then, because features in different granularity are mutually influenced, a bi-directional LSTM is proposed to model the interaction in the latent space. Finally, different encoders output the final prediction. The encoders and decoders consist of MLPs. In other words, the inter-granularity interaction network is proposed to learn the implicit links between trajectory and action, then the network utilizes the links to refine the predictions from intra-granularity and output more reasonable and accurate prediction results. For n-th individual at t time step, the decoder outputs Δ'_{act} and Δ'_{tra}, and $Q^c_{n,t} = X^c_{n,t} + \Delta'_{tra}$ and $Q^r_{n,t} = X^r_{n,t} + \Delta'_{act}$ are the final outputs representing the prediction for $t + 1$ time step.

3.3. Temporal Prediction

Each single-granularity LSTM of each person uses sequence-to-sequence architecture. Supposing the input video clip has T frames. After T time steps encoding, for n-th individual, we obtain the predicted value $Q^c_{n,T}$ and $Q^r_{n,T}$ from multi-granularity interaction prediction network at T time step, and $Q^c_{n,T}$ and $Q^r_{n,T}$ are the predictions for absolute coordinates of center points and relative coordinates of other joints of the n-th individual.

In the decode-stage, starting from $t = T$, we take $Q^c_{n,t}$ and $Q^r_{n,t}$ as the input to the different single-granularity LSTMs of n-th individual at $t + 1$ time step. And then the output $Q^c_{n,t+1}$ and $Q^r_{n,t+1}$ from multi-granularity interaction prediction network at $t + 1$ time step will be taken as the input to the single-granularity LSTMs at $t+2$ time step. After T_p time steps decoding, we obtain predictions including absolute coordinates of center points $C'_{n,t} = Q^c_{n,t}$ and relative coordinates of other joints $R'_{n,t} = Q^r_{n,t}$ for n-th individual at t time step, where $t = T + 1, T + 2, ..., T + T_p$ and $n = 1, 2, ..., N$. Then, we de-normalize the predictions and combine them together to reconstruct the skeleton data. Follows are the definitions of loss for training

3.4. Loss Definition

When modelling multi-granularity interaction, we focus more on loss in different granularity rather than absolute loss. To this end, we introduce multi-granularity loss as follows:

$$\begin{aligned}
\mathcal{L}_{tra} &= \sum_{n=1}^{N} \sum_{t=T+1}^{T+T_p} (C'_{n,t} - \hat{\mathcal{C}}_{n,t})^2, \\
\mathcal{L}_{act} &= \sum_{n=1}^{N} \sum_{t=T+1}^{T+T_p} (R'_{n,t} - \hat{\mathcal{R}}_{n,t})^2, \\
\mathcal{L} &= \mathcal{L}_{act} + \lambda \mathcal{L}_{tra} + \|W\|_2
\end{aligned} \tag{8}$$

where $C'_{n,t}$ is the predicted absolute coordinates of the center point, $R'_{n,t}$ is the predicted relative coordinates of other points, $\hat{\mathcal{C}}_{n,t}$ and $\hat{\mathcal{R}}_{n,t}$ are the corresponding ground truth. λ is used to balance the weights of action loss and trajectory loss.

4. Experiments

We perform extensive experiments on two challenging human interaction datasets, SBU Dataset [33] and Choi's New Dataset [9]. To quantify the impact of modeling multi-granularities, we not only compare the prediction results with related methods, but also with specific designed baselines. And further discussions are also provided.

4.1. Implementation Details

We implement our proposed method using Tensorflow [1]. In granularity features encoding, the hidden state size of trajectory LSTM and action LSTM are set as 32 and 128. In intra-granularity interaction encoding, the hidden units of encoders and decoders are easy to get in Figure.3. Besides, in inter-granularity interaction network, the hidden state size of bi-direction LSTM is set as 32. In order to avoid exploding gradient in LSTMs, we apply gradient clipping by 5. We trained our model in two steps (intra-granularity interaction, inter-granularity interaction). First,

Table 1. The average displacement error for SBU Dataset

Activity	*approching*				*kicking*				*punching*				*hugging*				*pushing*			
T_p	4	6	8	10	4	6	8	10	4	6	8	10	4	6	8	10	4	6	8	10
SP [19]	0.17	0.34	0.61	1.01	0.32	0.57	0.78	1.01	0.31	0.61	0.77	0.91	0.24	0.46	0.70	0.95	0.25	0.61	1.05	1.55
S-LSTM [2]	**0.11**	0.25	0.52	0.93	0.27	0.49	0.70	0.90	0.21	0.41	0.61	0.79	0.29	0.46	0.65	0.82	0.23	0.51	0.85	1.25
B-LSTM	0.14	0.28	0.51	0.83	0.23	0.39	0.62	0.89	0.19	0.36	0.54	0.73	0.22	0.35	0.49	0.68	0.19	0.33	0.47	0.61
SG-IN	0.12	0.23	0.38	0.57	**0.19**	**0.32**	0.51	0.71	**0.18**	0.35	0.50	0.67	0.19	0.36	0.53	0.69	**0.13**	0.24	0.37	0.52
O-Intra	0.15	0.31	0.55	0.83	0.38	0.65	0.96	1.31	0.26	0.46	0.67	0.92	0.24	0.47	0.64	0.86	0.17	0.34	0.58	0.89
MG-Concat	0.13	0.25	0.41	0.62	0.27	0.43	0.62	0.82	0.24	0.40	0.56	0.75	0.22	0.34	0.53	0.67	0.16	0.26	0.43	0.78
Ours	0.12	**0.17**	**0.23**	**0.31**	0.20	**0.32**	**0.48**	**0.63**	**0.18**	**0.32**	**0.46**	**0.60**	**0.17**	**0.32**	**0.44**	**0.65**	0.14	**0.23**	**0.32**	**0.47**

Table 2. The average displacement error for Choi's New Dataset

Activity	*gathering*				*queueing*				*walking together*				*chasing*			
T_p	4	6	8	10	4	6	8	10	4	6	8	10	4	6	8	10
SP [19]	0.30	0.56	0.89	1.16	0.35	0.67	0.98	1.33	0.27	0.37	0.62	0.87	0.32	0.46	0.68	0.97
S-LSTM [2]	0.27	0.42	0.56	0.77	0.32	0.48	0.65	0.83	0.26	0.35	0.43	0.53	0.28	0.39	0.53	0.67
B-LSTM	0.24	0.36	0.46	0.57	0.27	0.39	0.50	0.64	0.25	0.32	0.40	0.51	0.25	0.33	0.46	0.58
SG-IN	0.23	0.29	0.44	0.51	0.25	0.32	0.46	0.58	**0.23**	0.28	0.39	0.49	0.25	0.31	0.42	0.53
O-Intra	0.31	0.49	0.65	0.86	0.42	0.65	0.89	1.24	0.51	0.76	0.92	1.16	0.48	0.71	0.91	1.19
MG-Concat	0.21	0.33	0.45	0.71	0.34	0.56	0.73	0.85	0.32	0.43	0.62	0.81	0.33	0.48	0.67	0.83
Ours	**0.19**	**0.28**	**0.38**	**0.48**	**0.23**	**0.29**	**0.41**	**0.53**	**0.23**	**0.27**	**0.36**	**0.47**	**0.22**	**0.26**	**0.38**	**0.51**

we train the network only with inter-granularity interaction network with learning rate 0.001, and on this basis, we train the whole multi-granularity interaction network with learning rate 0.0005, on a single GPU (TITAN X) using Stochastic Gradient Descent algorithm.

In SBU dataset, we observe 6 frames (0.4*sec*) and predict 10 frames (0.67*sec*). In Choi's new dataset, we sample the video at 5HZ, and we observe 5 frames (1*sec*) and predict 10 frames (2*secs*). Then we analyze short-term and long-term prediction performance like [11, 12, 15, 19].

4.2. Baselines

To quantify the impact of our contributions, we design experiments from two aspects. First, to prove the effect of multi-person interaction and compare the ability of different structures for modeling interaction, we propose following single-granularity baselines. Note that the input to the network is the normalized skeleton data **without** using multi-granularity data generator.

- *Single-person method* (**SP**). We use the state-of-the-art single-person prediction method in [19] to model each individual in a group activity without considering interaction between each other.
- *Social LSTM* (**S-LSTM**). This baseline uses Social LSTM proposed in [2] to model interaction and predict skeleton data.
- *Bi-directional LSTM* (**B-LSTM**). Similar to S-LSTM, in this baseline, we use bi-directional lstm instead of social lstm.
- *Single-granularity interaction network* (**SG-IN**). This baseline uses the interaction network illustrated in Figure.3 to model multi-person interaction.

To illustrate the advantages of multi-granularity and to explore feasible interaction approach, following baselines are introduced. Note that input to the network is the outputs of multi-granularity data generator, *i.e.*, **with** multi-granularity data generator.

- *Only intra-granularity interaction* (**O-Intra**). Compared with our method in Figure.2, this baseline doesn't use inter-granularity interaction network.
- *Multi-granularity concat* (**MG-Concat**). In this baseline, features in different granularities are concatenated as input to SG-IN to model whole interaction.

4.3. Evaluation Metric

We report the prediction error with three metrics. First we define error as follows:

$$Err = \sum_{1}^{T} \sum_{1}^{N} (y_{i,t} - \hat{y_{i,t}})^2 \tag{9}$$

where N is the number of the individuals in the video, and T is the length of predicted frames.

Average displacement error: The error over all predicted skeleton data and the ground truth in a group activity. In this situation, $y_{i,t}$ is the predicted coordinates of all skeletal joints for i-th individual at t time step, and $\hat{y_{i,t}}$ is the ground truth.

Multi-granularity error: The error of features in different granularities including trajectory error and action error. For trajectory, $y_{i,t}$ is the predicted absolute coordinates of

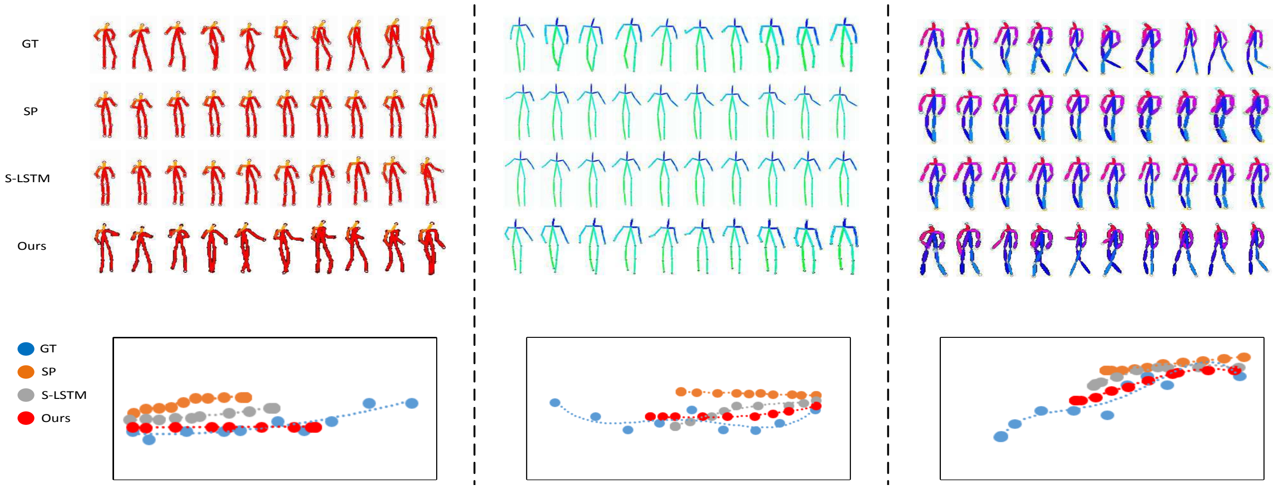

Figure 5. Predicted detailed actions and trajectories of each individual in a *gathering* activity using different methods

center joint. And for action, $y_{i,t}$ is the predicted relative coordinates of other joints

Skeleton joint error: We compute the error for each skeleton joint. For each joint, $y_{i,t}$ is the predicted coordinates of this joint.

4.4. SBU Dataset

SBU dataset[33] is an interaction dataset with two subjects. It contains about 300 sequences of 8 class interactions, including *approaching, departing, pushing, kicking, punching, exchanging objects, hugging, and shaking hands*. Except *departing*, there is significant interaction in the rest activities. Due to the lack of training data and the difference between training and testing data, we augment the data by flipping and rotation (we rotate all individuals as a whole according to z-coordinate based on 3D coordinate and project them to 2D). Besides we run 5-fold training and testing as suggested in [33] for each activity.

Table 1 illustrates the average displacement error of all methods. As shown in Table 1, SP method has high error because it ignores the interaction between individuals. And in single-granularity experiments, three baselines including S-LSTM, B-LSTM and SG-IN all outperform the SP baseline. More specifically, SG-IN performs best among the three, which proves SG-IN has greater capacity in modelling interaction. And in multi-granularity experiments, O-Intra and MG-Concat perform poorly, their error is even higher than using SP method in some activities. It emphasizes the importance of modelling interaction between features in different granularities and it is not a feasible way to concatenate them directly. Compared with all baselines, our method performs best in long-term prediction in all activities. Figure 6 shows an example of the *punching* activity using SP, S-LSTM and our method. As can be seen, in short-term prediction, the predictions among three methods are almost the same. However, with the increase of prediction time, the difference between the three methods is becoming more and more obvious, which has demonstrated the advantages of our method in long-term modelling group interaction.

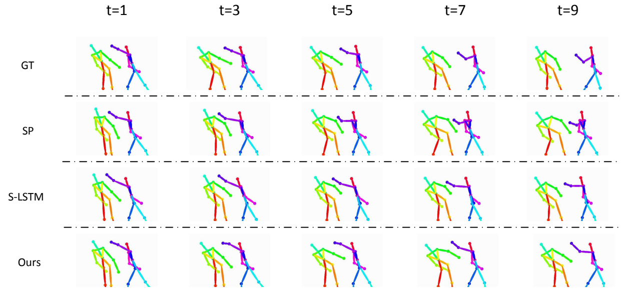

Figure 6. The results of *punching* activity in SBU dataset using three different methods

We also evaluate experimental results using multi-granularity error metric. The first row of Figure 7 shows an example of the *punching* activity. It has clearly demonstrated that our method is robust in predicting long-term interaction. With the increase of prediction time, the error of method increases slowest compared with other baselines. Meanwhile, it also shows that SG-IN outperforms S-LSTM and B-LSTM. Due to that most of skeleton joints are basically not moving in most of the interactions, we don't use the skeleton joint error metric to evaluate the results.

From these experiments, we demonstrate the effect of multi-person interaction and compare different interaction methods. Moreover, we validate the advantages of multi-granularity analysis and its feasible interaction approach.

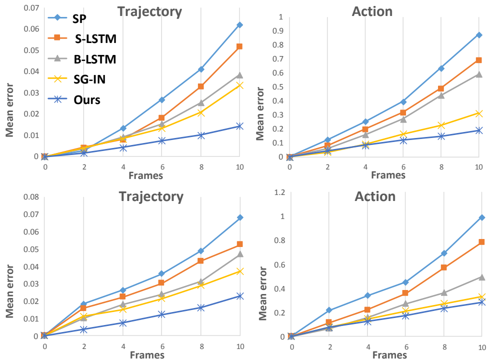

Figure 7. Multi-granularity errors including trajectory error and action error of *kicking* activity in SBU dataset and *approching* activity in Choi's New Dataset.

4.5. Choi's New Dataset

Choi's New Dataset [9] is composed of 32 video clips with 6 collective activities: *gathering, talking, dismissal, walking together, chasing and queueing*. Similarity, we take experiments on all activities except *talking, dismissal*. We use methods in [6] and [20] to estimate 3-dimension coordinates and augment data by rotation, flipping. We randomly divide the augmented data into 3 subjects and run 3-fold training and testing as suggested in [9].

For Choi's new dataset, the quantitative results of all baselines and our method are reported in Table 2. Different from SBU dataset, interactions in Choi's new dataset contain more than two subjects and individual trajectory changes are obvious, which make this dataset more challenging. And results show that our proposed method performs best in both short-term and long-term prediction. That proves that our multi-granularity interaction prediction method is more competent in group interaction prediction. It can also be proved from the last row of Figure 7. Figure 5 shows an example of *gathering* activity with three individuals. To better compare results from different methods, we display each person's action and trajectory separately. Using SP, the actions of all individuals convergence to a mean value quickly. Although, the results improves a little using S-LSTM, the actions are still not natural. And our method performs best and produces natural continuous action and trajectory.

In order to better compare with single-person method and other baselines, we track one person in the group interaction and compare mean errors of different skeleton joints. We split the joints into three parts: trunk, upper limbs and lower limbs. Figure 8 is the result of a test example from *gathering* in Choi's new dataset. It is obvious that the error of our method is lowest in most of joints, especially in the key points of movement, such as knees, ankles. Besides, the overall error of our method is the lowest.

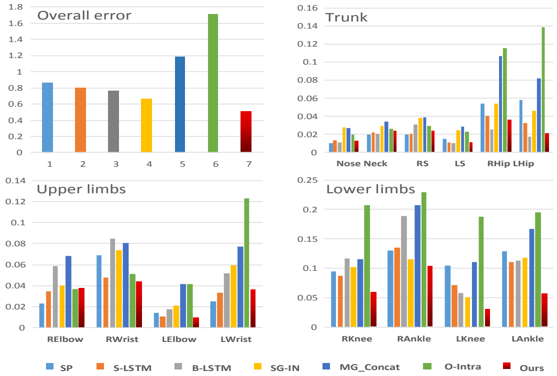

Figure 8. We track an individual in *approching* activity, and analyse the overall error and each joint's error using different methods.

4.6. Video Generation

Since we have predicted each person's skeleton, we use the method in [30] to generate videos based on predicted skeleton sequences. The following are two examples. We note that paired with our generated skeletons, the output videos are realistic.

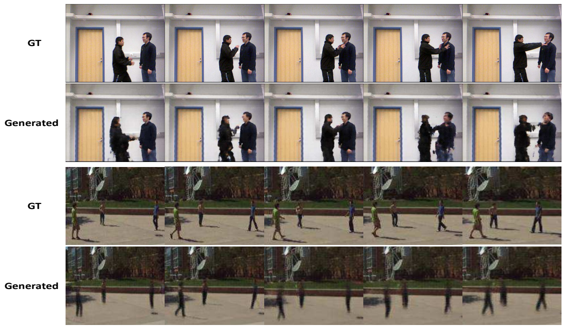

Figure 9. The generated video sequence based on predicted skeleton data

5. Conclusions

In this paper, we focus on group interaction prediction and propose a multi-granularity interaction network. We use intra-granularity interaction sub-network to capture interactions in the same granularity separately. And built on a bi-directional LSTM, the intgeraction network takes cross-granularity interaction into account and predicts long-term dynamic information of each individual in group activities. Results on two public datasets have validated the effectiveness and rationality of our method.

6. Acknowledgement

The work was supported by State Key Research and Development Program (2016YFB1001003). This work was partially supported by National Science Foundation of China (61502301,U1611461,61521062), China's Thousand Youth Talents Plan and 18DZ2270700.

References

[1] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *CVPR*, 2016.

[3] 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. *CVPR*, 2017.

[4] J. Bütepage, M. Black, D. Kragic, and H. Kjellström. Deep representation learning for human motion prediction and classification. *CVPR*, 2017.

[5] J. Bütepage, H. Kjellström, and D. Kragic. Anticipating many futures: Online human motion prediction and synthesis for human-robot collaboration. *arXiv preprint arXiv:1702.08212*, 2017.

[6] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *CVPR*, 2016.

[7] M. B. Chang, T. Ullman, A. Torralba, and J. B. Tenenbaum. A compositional object-based approach to learning physical dynamics. *Proceedings of the 5th International Conference on Learning Representations*, 2017.

[8] M. Chen, X. Jin, and D. Shen. Short text classification improved by learning multi-granularity topics. In *IJCAI*, 2011.

[9] W. Choi and S. Savarese. A unified framework for multi-target tracking and collective activity recognition. In *European Conference on Computer Vision*. Springer, 2012.

[10] W. Choi, K. Shahid, and S. Savarese. Learning context for collective activity recognition. In *CVPR*, 2011.

[11] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik. Recurrent network models for human dynamics. In *ICCV*. IEEE, 2015.

[12] P. Ghosh, J. Song, E. Aksan, and O. Hilliges. Learning human motion models for long-term predictions. *arXiv preprint arXiv:1704.02827*, 2017.

[13] M. S. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori. A hierarchical deep temporal model for group activity recognition. In *CVPR*, 2016.

[14] 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 Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014.

[15] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena. Structural-rnn: Deep learning on spatio-temporal graphs. In *CVPR*, 2016.

[16] T. Lan, L. Sigal, and G. Mori. Social roles in hierarchical models for human activity recognition. In *CVPR*, 2012.

[17] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori. Discriminative latent models for recognizing contextual group activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(8):1549–1562, 2012.

[18] A. M. Lehrmann, P. V. Gehler, and S. Nowozin. Efficient nonlinear markov models for human motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1314–1321, 2014.

[19] J. Martinez, M. J. Black, and J. Romero. On human motion prediction using recurrent neural networks. *CVPR*, 2017.

[20] J. Martinez, R. Hossain, J. Romero, and J. J. Little. A simple yet effective baseline for 3d human pose estimation. *ICCV*, 2017.

[21] V. Ramanathan, B. Yao, and L. Fei-Fei. Social role discovery in human events. In *CVPR*, 2013.

[22] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1010–1019, 2016.

[23] T. Shu, S. Todorovic, and S.-C. Zhu. Cern: Confidence-energy recurrent network for group activity recognition. *CVPR*, 2017.

[24] G. W. Taylor, G. E. Hinton, and S. T. Roweis. Modeling human motion using binary latent variables. In *Advances in neural information processing systems*, pages 1345–1352, 2007.

[25] P. Trautman and A. Krause. Unfreezing the robot: Navigation in dense, interacting crowds. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 797–803. IEEE, 2010.

[26] D. Wang, Z. Shen, J. Shao, W. Zhang, X. Xue, and Z. Zhang. Multiple granularity descriptors for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2399–2406, 2015.

[27] J. Wang, A. Hertzmann, and D. M. Blei. Gaussian process dynamical models. In *Advances in neural information processing systems*, pages 1441–1448, 2006.

[28] M. Wang, B. Ni, and X. Yang. Recurrent modeling of interaction context for collective activity recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[29] K. Yamaguchi, A. C. Berg, L. E. Ortiz, and T. L. Berg. Who are you with and where are you going? In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1345–1352. IEEE, 2011.

[30] Y. Yan, J. Xu, B. Ni, and X. Yang. Skeleton-aided articulated motion generation. *arXiv preprint arXiv:1707.01058*, 2017.

[31] R. Yang, B. Ni, C. Ma, Y. Xu, and X. Yang. Video segmentation via multiple granularity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[32] R. Yin, Q. Wang, P. Li, R. Li, and B. Wang. Multi-granularity chinese word embedding. In *EMNLP*, pages 981–986, 2016.

[33] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 28–35. IEEE, 2012.

[34] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, X. Xie, et al. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *AAAI*, 2016.