Participation-Contributed Temporal Dynamic Model for Group Activity Recognition

Conference Paper · October 2018

DOI: 10.1145/3240508.3240572

5 authors, including:

Rui Yan
Nanjing University of Science and Technology
28 PUBLICATIONS **514** CITATIONS
SEE PROFILE

Jinhui Tang
Nanjing University of Science and Technology
329 PUBLICATIONS **18,001** CITATIONS
SEE PROFILE

Xiangbo Shu
Nanjing University of Science and Technology
72 PUBLICATIONS **2,718** CITATIONS
SEE PROFILE

Zechao Li
Chinese Academy of Sciences
88 PUBLICATIONS **3,734** CITATIONS
SEE PROFILE

Participation-Contributed Temporal Dynamic Model for Group Activity Recognition

Rui Yan
School of Computer Science and Engineering, Nanjing University of Science and Technology
ruiyan@njust.edu.cn

Jinhui Tang*
School of Computer Science and Engineering, Nanjing University of Science and Technology
jinhuitang@njust.edu.cn

Xiangbo Shu
School of Computer Science and Engineering, Nanjing University of Science and Technology
shuxb@njust.edu.cn

Zechao Li
School of Computer Science and Engineering, Nanjing University of Science and Technology
zechao.li@njust.edu.cn

Qi Tian
Huawei Noah's Ark Lab
Department of Computer Science, University of Texas at San Antonio
tian.qi1@huawei.com

ABSTRACT

Group activity recognition, a challenging task that a number of individuals occur in the scene of activity while only a small subset of them participate in, has received increasing attentions. However, most of the previous methods model all the individuals' actions equivalently while ignoring a fact that not all of them are contributed to the discrimination of group activity. That is to say, only a small number of key actors (participants) play important roles in the whole group activity. Inspired by this, we explore a new "One to Key" idea to progressively aggregate temporal dynamics of key actors with different participation degrees over time from each person. Here, we focus on two types of key actors in the whole activity, who steadily move in the whole process (long moving time) or intensely move (but closely related to the group activity) at a significant moment. Based on this, we propose a novel Participation-Contributed Temporal Dynamic Model (PC-TDM) to recognize group activity, which mainly consists of a "One" network and a "One to Key" network. Specifically, "One" network aims at modeling the individual dynamic of each person. "One to Key" network feeds the outputs from the "One" network into a Bidirectional LSTM (Bi-LSTM) according to the order of individual's moving time. Subsequently, each output state of Bi-LSTM weighted by a trainable time-varying attention factor is aggregated by going through LSTM one-by-one. Experimental results on two benchmarks demonstrate that the proposed method improves group activity recognition performance compared to the state-of-the-arts.

CCS CONCEPTS

• **Computing methodologies → Activity recognition and understanding**; *Hierarchical representations*;

KEYWORDS

Group activity recognition; long short term memory; video analysis

ACM Reference Format:
Rui Yan, Jinhui Tang, Xiangbo Shu, Zechao Li, and Qi Tian. 2018. Participation-Contributed Temporal Dynamic Model for Group Activity Recognition. In *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240572

1 INTRODUCTION

Activity recognition which aims to enable computer to understand the actions appeared in a video clip has received a great deal of research attention in computer vision and multimedia communities [12, 14, 29, 38, 41]. According to the number of participants, human activities can be mainly divided into three categories: single-person action [12, 33], human interaction [32, 39] and group activity [17, 30]. Previous works [12, 33, 38] paid more attention to single-person action recognition, and have made good progress. Besides of single-person action, a real scenario may contains more human interactions (e.g.,"handshaking") and multi-person activity (e.g., "Queueing", "Playing together"). In a scene of human interaction, there are at least two persons who are concurrently interacting with each other. In a scene of group activity, the activity depicts a more complex scene/event involving single-person action and various other interactions (e.g. group-person and group-group interaction). Generally, compared with single-person action recognition and human interaction recognition, group activity recognition is a more challenging task [17, 39].

Generally, existing solutions for group activity recognition can be summarized as two key steps: 1) understanding individual action from the motion descriptor of each person; and 2) inferring the class label of the group activity from a collection of individual motions. In the first step, some works [17, 30, 39] are proposed to learn

*Corresponding author.

MM '18, October 22–26, 2018, Seoul, Republic of Korea
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240572

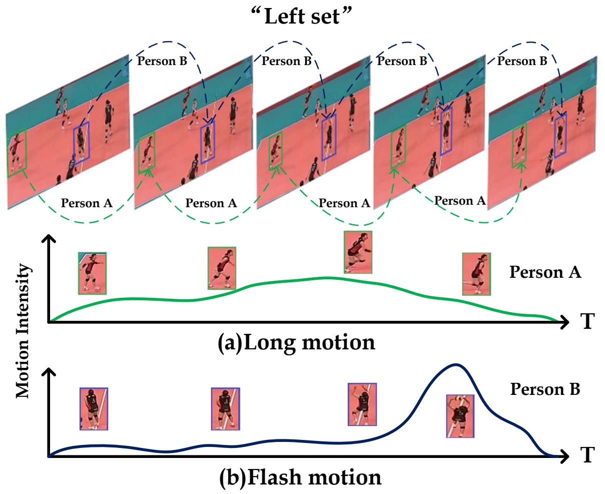


Figure 1: Illustration of key participants in a “Left set” activity of volleyball match. Two curves are plotted to reflect Person A and Person B's motion intensity and moving time, respectively. Person A keeps moving during the whole process of this video clip, while person B hits the ball (related to “left set” activity) with a intensive motion at a moment. These two types of motions are called long motion and flash motion respectively. Obviously, the participants with these two types of temporal motions are more related to the “left set” activity of volleyball match. Thus such participants are called key actors in this activity.

hierarchical feature representation with Convolutional Neural Networks (CNN) for better understanding individual's action, which has achieved better performance than the hand-crafted feature based methods [1, 7, 23]. In the second step, some recent methods attempt to encode the high order relationship among persons in the scene by graphical structures [2, 10] or Recurrent Neural Networks (RNN) [39]. However, most of these works ignore such inherent characteristic of group activity that not all individuals' actions are contributed to group activity equivalently.

In a group activity, Deng et al. [10] validated that some individuals are irrelevant to the whole group activity [24], who can be outliers compared to the group activity. That is to say, only a small number of key actors (participants) play important roles in the whole group activity. Therefore, it is essential to discover these key actors for recognizing the group activity effectively. The fundamental problem becomes: *Who are the key actors?* We investigate that key actors should have steady motions during the whole process or remarkable motions at a moment. To better explain this, we give an example of key actors' motion in volleyball match, as shown in Figure 1. In the “left set” scene of volleyball match, person A moves across the court and participates in the activity of “left set” during the whole video, which is called long motion. And, person B only has a sudden motion (hit the ball) at a significant moment. Although this sudden motion is short, it is closely related to the “left set” activity. In this paper, this type of motion is called flash motion. Both types of key actors can provide crucial clues to understand the activity “left set”.

To model the important dynamics of key actors while avoiding the irrelevant dynamics of outlier persons, we propose a novel Participation-Contributed Temporal Dynamic Model (**PC-TDM**) for group activity recognition. The framework of PC-TDM on a volleyball match is shown in Figure 2. First, we extract spatial features of each person on the detected and tracked bounding boxes by employing a pre-trained CNN model. Second, we take the spatial features of each person as the input of a special **“One” Network**, to model individual dynamics of each person over time. Third, the output states of “One” Network are fed into the proposed **“One to Key” Network (OKN)** for aggregating to the discriminative motion information of group activity scene, by attending to the key actors while avoiding the irrelevant outlier persons. More specifically, the **Interaction Bi-LSTM** models the individuals' interactions in accordance with the order of individual's long motions throughout the whole activity process. Then the **Aggregation LSTM** aims to aggregate latent output states of Interaction Bi-LSTM with the trainable attention weights progressively. Here, an attention weight describes the intensity of an individual's flash motion, which is varying with time. Finally, the concatenated states at each time step input to a softmax layer, and the averaged softmax score on all time steps is the prediction probability vector of group activity class.

Overall, the contributions of this work can be summarized as follows:

- To recognize group activity with a subset of participants, we propose a novel Participation-Contributed Temporal Dynamic Model to hierarchically learn discriminative feature descriptors from each person to key participants. And then we conduct experiments to illustrate the superiority of the proposed method by comparing with the state-of-the-art methods.
- As the first time to consider the participation degrees of all persons for group activity recognition, the proposed **“One to Key” Network** can progressively aggregate the dynamics of key participants with different long motions or flash motions one-by-one, while automatically avoiding the irrelevant motions of outlier persons. This can be flexibly embedded into the other network architectures.

In experiments, we evaluate the performance of the proposed PC-TDM on two widely-used datasets (e.g., Volleyball Dataset and Collective Activity Dataset) by comparing with the state-of-the-art methods and several baselines.

2 RELATED WORKS

For the past few years, group activity recognition [3, 11, 16, 22, 40] has developed into an attractive topic in computer vision and multimedia areas. In the early stages, many works employ hand-craft feature to represent individuals' motions in space and/or time domain[7, 8, 24, 27]. In particular, Choi et al. [7] designed a local spatio-temporal descriptor to capture the spatial distribution of pedestrians effectively over time, and then release a collective activity dataset. Compared with traditional machine learning methods [7, 24, 31], Deep Neural Networks (DNNs) have shown excellent performance in a variety of computer vision tasks [18, 26, 36, 37]. Thus many DNN-based methods for group activity recognition have sprung up. Ibrahim et al. [17] proposed a hierarchical model

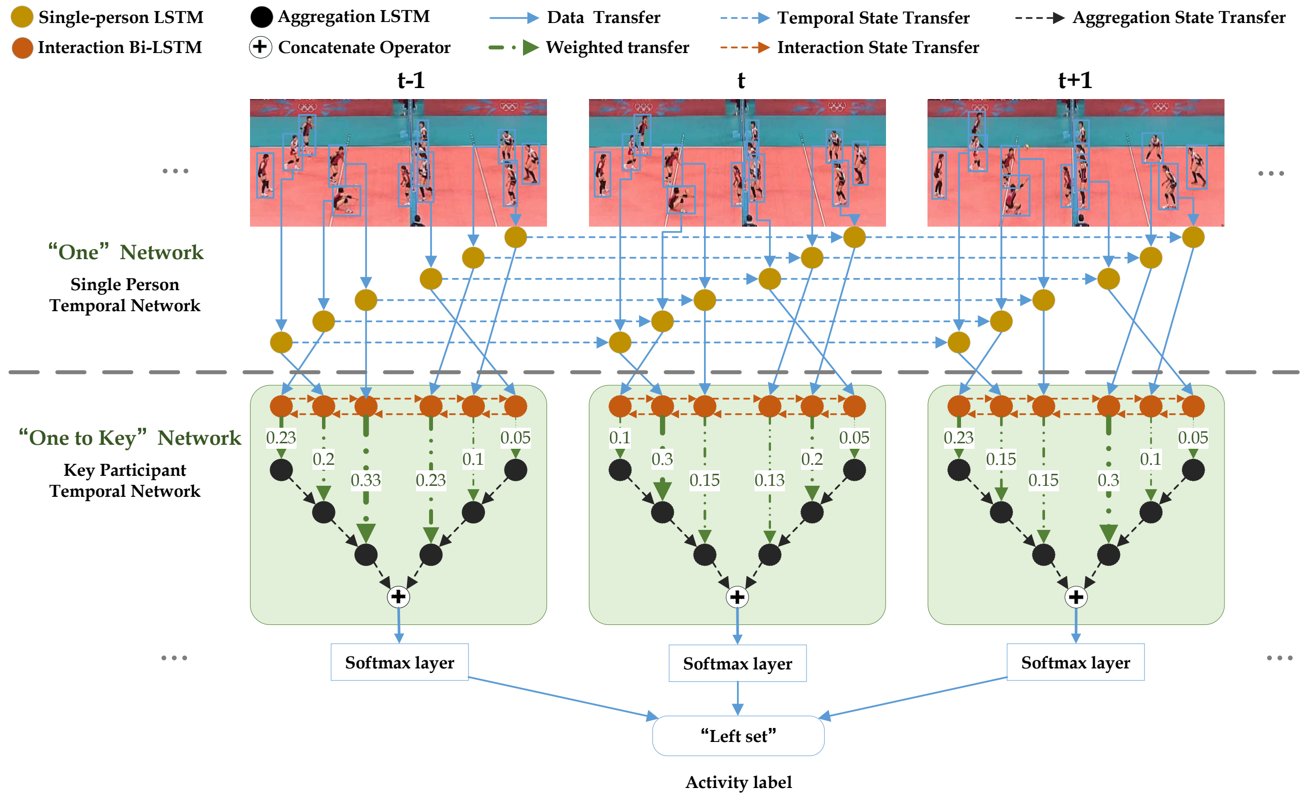


Figure 2: Framework of the proposed PC-TDM for recognizing group activity on a volleyball match. First, we extract their spatial CNN features of each individual at each time step. Second, we feed these spatial CNN features into Single-Person LSTM for capturing the individual dynamics. Third, the “One to Key” Network consisting of two layers of LSTM (i.e., Interaction Bi-LSTM and Aggregation LSTM) is designed to aggregate spatio-temporal features with attending to key actors. Specifically, we feed the individual dynamics of each person into the Interaction Bi-LSTM in accordance with the order of individuals moving time throughout the whole activity process. Then we integrate all hidden states from Interaction Bi-LSTM with the time-varying attention weights by an Aggregation LSTM, and concatenate the aggregated states of two sides as an input of a softmax layer at each time step. Finally, we average softmax scores from each time step as the final prediction probability vector for the group activity recognition.

with serval LSTM layers to recognize group activity ranging from low-level to high-level dynamics. Similarly, Wang et al. [39] extended a RNN-based hierarchical framework to handle three level interactions (single human dynamics, within group human interaction and group to group interaction) with a high order context modeling scheme.

Recently, some works [30, 32] argued that the individual actions are related to each other, and proposed to model the interaction-related dynamics over time. For example, Shu et al. [30] proposed a Confidence-Energy Recurrent Network to integrate the confidences from two types of predictions (individual action prediction and human interaction prediction) to energy layer in inferring the class label of event. Shu et al. [32] designed a Concurrence-Aware Long Short-Term Sub-Memories to explore the long-term inter-related dynamics among interacting individuals, rather than the individual dynamics of each person.

In addition to the temporal sequence, the complex spatial structure among persons also existing in a group activity. Deng et al. [10] proposed a Structure Inference Machine with RNN to iteratively update the graphic model, and to reason about which people in a scene are interacting and infer the label of group activity. Furthermore, Jain et al. [2] extended the traditional spatio-temporal structure data into a Spatio-Temporal Graph model within the units of RNN to effectively infer the action/relation between humans and objects.

However, most of these works did not realize that only a small number of persons’ actions are closely related to the whole group activity. Although Yan et al. [42] predicted human interaction via relative interacting region, and Ramanathan et al. [28] proposed to detect the event by attending to the key persons responsible for the event in a multi-person video clip. Contrast to [28, 42], this work aims to capture the key participants with at least one of the key characteristics (i.e., long motion and flash motion) for group activity recognition. Unlike [28], we employ optical flow to measure the intensity of individual’s long motion, and learn a attention factor to describe the intensity of individual’s flash motion over time.

3 THE PROPOSED METHOD

3.1 Preliminary

In a video clip with T frames, it describes a specific group activity including K persons. Let Z_t denotes the representation of the t-th frame, and X_t^k denotes the representation of the k-th person in the t-th frame, where $t \in \{1, 2, \cdots, T\}$, and $k \in \{1, 2, \cdots, K\}$. We observe the actions of all the people from time step 1 to T, and recognize what a group of people are doing in this short video, which is called group activity recognition in this paper.

Actually, a group activity contains multiple individuals who are doing various actions, but not all of them are engaged in the activity. That is to say, in the group activity recognition task, the class (e.g., "Moving", "Talk") of group activity is related to a smaller subset of individuals, i.e., key actors. Inspired by this, we consider to infer the class label of group activity by mainly capturing the effective motion information of these key actors from all person in the activity scene. Intuitively, the key actors should have the following characteristics: steadily moving in the whole activity process or intensively moving at a moment, as shown in Figure 1. In this paper, they are called long motion and flash motion respectively, which are defined as follows,

- **Long motion** is performed by an actor who steadily moves throughout the whole activity process, such as "moving on the court". Thus, in the whole activity process, the person who has the longest moving time should have the longest motion. To measure the time of long motion of individual, we adopt to calculate the mean value of the optical flow corresponding to such person between all two consecutive frames.
- **Flash motion** is performed by an actor who intensely moves at a significant moment, such as "spiking the ball", which is closely related to the group activity. Obviously, the flash motion provides the crucial clue to represent the semantic of the group activity in a video clip. Since flash motion often happens at a significant moment in a video clip, we consider to adopt the attention strategy to force it over time.

To model the important dynamics of key actors with long motion or flash motion well, a novel **Participation-Contributed Temporal Dynamic Model** (PC-TDM) is proposed in this work. The proposed PC-TDM contains two modules, i.e., **"One" Network**, and **"One to Key" Network** (OKN), as shown in Figure 2. In "One" Network, Single-Person LSTM models the individual dynamics from the spatial CNN features of each person. In "One to Key" Network (OKN), its goal is to aggregate the personal spatio-temporal features with attending to the key actors, while avoiding the outlier persons. Specifically, Interaction Bi-LSTM models interaction dynamics among individuals in accordance with the order of their long motions throughout the whole activity process. And Aggregation LSTM progressively aggregates the latent states from Interaction Bi-LSTM with trainable time-varying attention weights. These details are described in the following sections.

3.2 "One" Network: Single Person Temporal Network

Generally, a group activity includes a number of persons, and most of them participate in this group activity. As we know, it is easy to distinguish between "Moving" and "Waiting", depending on the individual actions of most of person. Hence, it is primary to learn individual temporal representation of each person for recognizing the group activity. There, similar to [17, 32, 39], we build a "One" Network to model individual dynamics of each person well. Specifically, we extract CNN features from each person's tracklet at each time step as the static representations of individuals. And then we leverage a long short-term memory (LSTM) (called Single-Person LSTM in this paper) to model the individual dynamics from the static representations of individuals.

Formally, we denote the sequence of CNN features of one individual by $X = \{x_1, x_2, ..., x_T\}$, where x_t is the spatial CNN features at time step t extracted from a pre-trained CNN model. The input gate i_t, forget gate f_t , output gate o_t, and input modulation gate g_t , memory cell c_t of **Single-Person LSTM** are defined as follows,

$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i); \tag{1}$$

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f); \tag{2}$$

$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o); \tag{3}$$

$$g_t = \sigma(W_{gx}x_t + W_{gh}h_{t-1} + b_g); \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t; \tag{5}$$

$$h_t = o_t \odot \varphi(c_t), \tag{6}$$

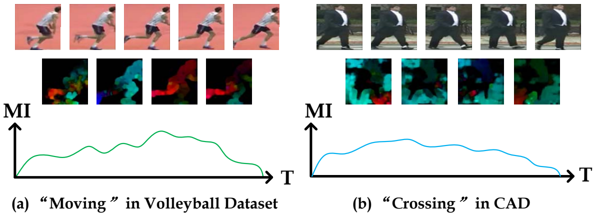
where $\sigma(*)$ is a sigmoid function; W_{*x} and W_{*h} are the weight matrices; b_* is the bias vector; $\odot$ denotes the element-wise product; as well as h_t is the hidden state which contains the dynamics of that person at time step t.

3.3 "One to Key" Network: Key Participant Temporal Network

Since "One" Network modeling the individual dynamics of each person equivalently and independently, it ignores two facts that 1) the actions of some outlier persons in the activity scene are irrelevant to this group activity; and 2) the actions of some persons are related to each other.

To this end, we propose an "One to Key" Network (OKN) following "One" Network to sequentially model the dynamics of key participants with **long motions** and **flash motions** from the individual dynamics of each person. Specifically, we firstly employ Bi-LSTM to model the individuals' interactions in accordance with the order of individual's long motions throughout the whole activity process, and then design a Aggregation LSTM to progressively aggregate latent output states of Bi-LSTM with trainable attention weights. Details are given as follows.

3.3.1 ***Modeling for Long Motion.*** Long motion is often performed by a participant who has continuous motion throughout the whole activity process. The longer the moving time of one person has, the more important role she/he plays. To measure the moving time of long motion in the entire video clip, we attempt at measuring mean motion intensity for each person by stacking the



(a) "Moving" in Volleyball Dataset
(b) "Crossing" in CAD

Figure 3: Examples of Motion Intensity (MI). Given the tracklets of a person on all frames, we calculate their optical flows between each two consecutive frames, then draw her/his motion intensity over time.

optical flow [5, 12, 33] images and calculating the mean value of them, as illustrated in Figure 3.

Mathematically, given a video clip of T frames and the resolution of each frame is $w*h$, we denote the horizontal and vertical displacement vectors at point (u, v) $(u = 1, 2, ..., w$, and $v = 1, 2, ..., h)$ in frame t by $d_t^x(u, v)$ and $d_t^y(u, v)$, respectively. Firstly, we stack flow vectors $d_t^x(u, v)$ and $d_t^y(u, v)$ of T consecutive frames as follow:

$$SF^k(u, v, 2i - 1) = d_t^x(u, v); \tag{7}$$

$$SF^k(u, v, 2i) = d_t^y(u, v), \tag{8}$$

where $i = 1, 2, \cdots, T$. Then we obtain $SF^k(u, v, c)$ $(c = 1, \cdots, 2T)$ that encodes the motion of k-th person at point point (u, v) over a sequence of T frames, and define the intensity of long motion w.r.t. k-th person as:

$$MI_t^k = \frac{\sum_{u=0}^{w}\sum_{v=0}^{h}\sum_{c=2t-1}^{2t}|SF_k(u, v, c)|}{w * h}; \tag{9}$$

$$MI^k = (\sum_{t=1}^{T} MI_k^t)/T, \tag{10}$$

where MI_k^t is the motion intensity of k-th person at time step t, and MI_k is the motion intensity of k-th person throughout the whole process. Obviously, if MI_k of one person is large, it is indicated that this person participates in the activity frequently over time.

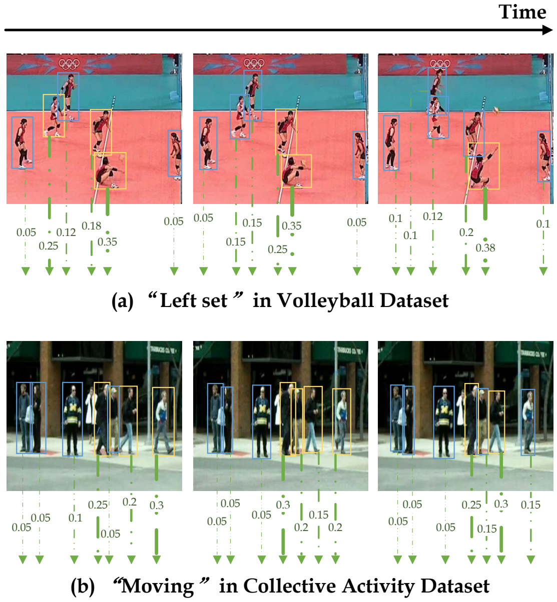
To model interaction-related dynamics among persons via RNN, some works [17, 39] ordered an interaction sequence of person by roughly using spatial positions of all persons. This ignores a fact that some closer persons are not related sometimes. Obviously, one person who keeps moving (e.g., "Moving", "Jumping") has a large amount of time to interact with other persons over many time steps. Therefore, the moving person with long moving time should be early modeled by LSTM. Formally, we rank individual features of each person via the values of MI_k in descending order, which is taken as an interaction sequence for inputting to LSTM. Considering that the interaction between two persons is bi-directional, we utilize a new Interaction Bi-LSTM rather than the traditional LSTM to model such interaction sequence. At time step t, the **Interaction Bi-LSTM** unit computes the forward hidden sequence $\{\overrightarrow{h}_t^1, \overrightarrow{h}_t^2, \cdots, \overrightarrow{h}_t^k, \cdots, \overrightarrow{h}_t^K\}$ and the backward hidden sequence $\{\overleftarrow{h}_t^1, \overleftarrow{h}_t^2, \cdots, \overleftarrow{h}_t^k, \cdots, \overleftarrow{h}_t^K\}$ by iterating K persons from $k = K \rightarrow 1$ and $k = 1 \rightarrow K$ respectively. Then the output sequence $\{\tilde{h}_t^k, \tilde{h}_t^2, \cdots, \tilde{h}_t^k, \cdots, \tilde{h}_t^K\}$ can be computed as follows,

$$\overrightarrow{h}_t^k = \mathcal{H}(W_{x\overrightarrow{h}}\tilde{x}_t^k + W_{\overrightarrow{h}\overrightarrow{h}}\overrightarrow{h}_t^{k-1} + b_{\overrightarrow{h}}); \tag{11}$$

$$\overleftarrow{h}_t^k = \mathcal{H}(W_{x\overleftarrow{h}}\tilde{x}_t^k + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_t^{k+1} + b_{\overleftarrow{h}}); \tag{12}$$

$$\tilde{h}_t^k = \overrightarrow{h}_t^k \diamond \overleftarrow{h}_t^k, k = 1, 2, \cdots, K, \tag{13}$$

where $\mathcal{H}$ is implemented by Eq. (1)-(6), and $\diamond$ denotes the pooling operation. Contrast to concatenating the forward and backward hidden sequence in typical Bi-LSTM, we construct the final sequence representation $\tilde{h}$ by pooling $\overleftarrow{h}_t^k$ and $\overrightarrow{h}_t^k$. This not only eliminates the redundant information, but also reduces the computational overhead of model.



(a) "Left set" in Volleyball Dataset
(b) "Moving" in Collective Activity Dataset

Figure 4: Different intensive flash motions of participants in a moment. The value of attention factor describes the intensive degree of flash motion.

3.3.2 *Modeling for Flash Motion*.

Besides on long motion, some persons are not steadily moving in the whole activity, while they have intensive motions in a significant moment, namely flash motion. These motions also provide important discriminative information for recognizing groups activity by RNN model. Taking the "left set" activity in volleyball match as an example, as shown in Figure 4(a), several persons (in yellow bounding boxes) around the volleyball participate in activity with more intensive flash motion at a moment. Their motions are closely related to "left set" activity provide the crucial information to understand this activity.

Since the flash motion is varying over time, we consider to assign different attention factors to force the intensity of flash motion of key participants over time. One straight way is that we can compute the attention factor of one person at one time step based on the value of the optical flow between two consecutive frames. However, some flash motions happening at a significant moment may be not related to group activity, such as "one person suddenly tumbles in a set activity of volley match", "two persons are colliding in a walking activity", etc.

In this work, we build a **Aggregation LSTM** to learn the attention factor for each person via her/his prediction probability of person-level action, and then progressively aggregates latent output states of Interaction Bi-LSTM. Here, if the person-level action one person is more consistent with the group-level activity, the corresponding learned attention factor will be larger, vice versa. To improve the model capabilities, we adopt the strategy used in [17] to aggregate the individuals' information into one representation. For the volleyball dataset, the two branches correspond to two teams of players respectively. The features from two branches are concatenated to form one representation to avoid the confusion of "left" and "right". Specifically, we split the whole group activity within K persons into N_g sub-groups for recognition, where $g = 1, 2, \cdots, N_g$. The start-index S_g and end-index E_g of persons in the g-th sub-group can be re-expressed as follows,

$$S_g = (g-1) * K/N_g + 1; \tag{14}$$

$$E_g = g * K/N_g. \tag{15}$$

In g-th sub-group of a video clip, for the k-th person, we capture the intensity of her/his flash motion by learning an attention factor γ_t^k to control her/his state $\tilde{h}_t^k$ from Interaction Bi-LSTM at time step t:

$$\hat{X}_{tg} = [\gamma_t^{S_g}\tilde{h}_t^{S_g}; \gamma_t^{S_g+1}\tilde{h}_t^{S_g+1}; \cdots; \gamma_t^{E_g}\tilde{h}_t^{E_g}]; \tag{16}$$

$$\gamma_t^k = \frac{\exp(e_t^k)}{\sum_{i=1}^{E_g-S_g+1}\exp(e_t^i)}, \tag{17}$$

where $e_t^k = Relu(W_{he}\tilde{h}_t^k + b_e)$, $k \in \{S_g, S_g+1, ..., E_g\}$; W_{he} is the weight matrix, b_e is the bias vector, and $exp(*)$ is the exponential function. Then we achieve the latent representation $\hat{X}_{tg}$ for each person in g-th group at time step t. So far, an Aggregation LSTM unit at time step t can be simply expressed as follows,

$$\hat{h}_{tg}^k = \textbf{Aggregation\_LSTM}(\hat{h}_{tg}^{k-1}, \hat{X}_{tg}^k); \tag{18}$$

$$Z_{tg} = \hat{h}_{tg}^{E_g}, \tag{19}$$

where Z_{tg} is the representation of g-th sub-group at time step t. Next, we get the activity representation by concatenating the features from all N_g sub-groups:

$$Z_t = Z_{t1} \oplus Z_{t2} \cdots \oplus Z_{tN_g}. \tag{20}$$

Finally, we feed the concatenated Z_t at each time step into a softmax layer, and average them over frames as the final prediction vector of group activity class.

4 EXPERIMENTS

In experiments, we evaluate the performance of proposed PC-TDM on two benchmarks by comparing with the state-of-the-art methods and several baselines.

4.1 Datasets

Two benchmarks used in experiments are introduced as follows,

- **Volleyball Dataset [17].** It is a new sport dataset collected from publicly available YouTube volleyball videos, consists of 55 videos with 4830 annotated frames. For one frame, the location of each player is given and labeled with one of the action classes (e.g. "Waiting", "Setting", "Digging", "Failing", "Spiking", "Blocking", "Jumping", "Moving" and "Standing"), and one of the group activity classes (e.g. "Left pass", "Right pass", "Left set", "Right set", "Left spike", "Right spike", "Left winpoint" and "Right winpoint") is labeled to this frame. For comparison with state-of-the-art methods and baselines B1-B4, we use the following performance metrics: 1) multi-class classification accuracy (MCA), and 2) mean per-class accuracy(MPCA). Our split of training and testing sets is the same as in [17].
- **Collective Activity Dataset (CAD) [7].** It contains 44 video clips collected by a low resolution hand-held camera. Each person labeled from five action labels (i.e., "Crossing", "Waiting", "Queuing", "Walking" and "Talking") and eight pose labels (not used in our work). A scene is assigned with the label of group activity based on what the majority of people are doing in the scene. We follow the train/test split provided by [13], and use the tracklet data provided in [6]. Following the experimental setting in [39], we merge class "Walking" and "Crossing" as "Moving" and report the Mean Per Class Accuracy (MPCA) due to the imbalanced test set.

4.2 Baselines

In experiments, four baselines are set as follows,

B1 Single Person Classification: In this baseline, we deploy the pre-trained AlexNet CNN to extract fc7 features on bounding boxes corresponding to each person, and max-pooled them to a single representation at each time step. Finally, we use these pooled features of individuals to train a softmax classifier. This baseline is designed to illustrate the importance of deep features.

B2 PC-TDM without OKN: This baseline is a variant of the proposed model which omitting OKN, and recognizes activity depend on max-pooling over all personal spatio-temporal features for each frame directly. This baseline aims to illustrate the importance of temporal dynamic.

B3 PC-TDM without Long Motion: This baseline is a degraded version of the proposed model that orders the personal features by roughly using spatial positions of all persons (i.e., from left to right in a frame). This baseline can illustrate the effectiveness of long motion.

B4 PC-TDM without Flash Motion: This baseline is a degraded version of the proposed model that discharges the Aggregation LSTM. The hidden states output from Interaction Bi-LSTM are max-pooled into a single representation at each time step, and input to the softmax classifier. This baseline can illustrate the effectiveness of flash motion.

4.3 Implementation Details

The input to our model are a set of bounding boxes (tracklets) around each person tracked over T frames by the object tracker [9], implemented in the Dlib library [19]. Our proposed framework is

Table 1: Comparison of different methods on Volleyball Dataset. [4, 30] do not provide specific accuracy of per-class or the mean per-class accuracy (MPCA). [25] ignores the classes of "Left winpoint" and "Right winpoint".

Methods	*L_pass*	*R_pass*	*L_set*	*R_set*	*L_spike*	*R_spike*	*L_winpoint*	*R_winpoint*	MCA	MPCA
Ibrahim *et al.* [17]	77.9	81.4	84.5	68.8	89.4	85.6	88.2	87.4	81.9	82.9
Shu *et al.* [30]	-	-	-	-	-	-	-	-	83.3	83.6
Li *et al.* [25]*	55.8	69.1	67.3	52.1	82.1	79.2	-	-	66.9	67.6
Biswas *et al.* [4]	-	-	-	-	-	-	-	-	83.5	-
B1	73.0	73.8	83.3	70.8	86.0	87.9	74.5	47.1	76.2	74.6
B2	81.9	77.1	85.7	74.0	88.3	88.4	79.4	47.1	79.7	77.7
B3	82.7	**88.6**	93.5	74.5	91.6	89.6	**92.2**	75.9	86.2	86.1
B4	**89.8**	83.3	**94.1**	**80.2**	86.6	**92.5**	83.3	73.6	86.3	85.4
PC-TDM	85.8	88.1	90.5	**80.2**	**92.2**	87.9	89.2	**90.8**	**87.7**	**88.1**

adaptive to various complex networks (e.g. VGG [34], ResNet [15] and GoogLeNet [35]) for feature representation in individuals' actions recognition stage. For fair comparison to [17, 39], we employ the pre-trained AlexNet model [21] to extract CNN features on bounding boxes corresponding to each person. Similar to [17], we train the propose PC-TDM in a stage-wise manner. Specifically, we firstly train "One" Network consisting of CNN and LSTM layer in an end-to-end manner to recognize individuals' actions. And then, the concatenation of spatial CNN and temporal features output from "One" Network are passed to the "One to Key" Network for group activity recognition. All the codes of experiments are implemented with Pytorch toolbox on a NVIDIA Tesla K20 GPU. We use the Adam algorithm [20] with the learning rate of 0.001 for all networks to minimize the cost function, and the learning rate is decreased to 1/10 of the original value after every ten epochs.

In experiments on Volleyball Dataset, 10 time steps and 3000 hidden nodes are used for the Single-Person LSTM and a softmax layer is deployed for the classification in the "One" Network. The number of sub-group is set to $N_g = 2$. In "One to Key" Network (OKN), Interaction Bi-LSTM has a six time steps (there are six individuals in one sub-group) and 1000 nodes; and Aggregation LSTM has six time steps and 1000 nodes.

In experiments on Collective Activity Dataset, 10 time steps and 3000 hidden nodes are used for the Single-Person LSTM and a softmax layer is deployed for the classification in the "One" Network. The number of sub-group is set to $N_g = 1$, namely we do not need to divide group. In "One to Key" Network (OKN), Interaction Bi-LSTM has a five time steps (there are six individuals in one sub-group) and 1000 nodes; and Aggregation LSTM has five time steps and 1000 nodes. Since the number of individuals in this dataset is varying from 1 to 12. We select five effective persons for each frame and regard them as an entire group. If the number of persons is less than five, we take a full-zero matrix as the tracklets of new person.

4.4 Results on Volleyball Dataset

Comparison with baselines. Table 1 shows the recognition accuracy of the proposed PC-TDM compared with the baselines. As shown in this table, the proposed PC-TDM achieves the best MCA and MPCA at the same time compared to all baseline methods. The results of B1 and B2 illustrate the importance of deep features and temporal dynamics respectively. Compared to B3 and B4, our PC-TDM considering the key participants with long motion or flash motion obtains better performance.

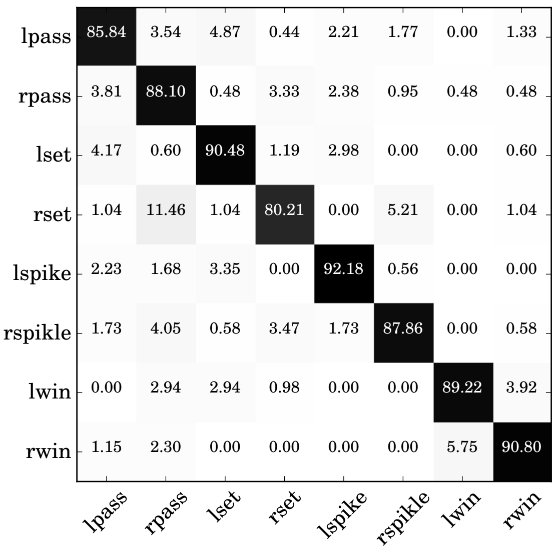


Figure 5: Confusion matrix of the proposed PC-TDM on Volleyball Dataset.

Comparison with state-of-the-art methods. There are only a little of literatures (e.g. Ibrahim et al. [17], Shu et al. [30], Li et al. [25], Sovan et al. [4]) referring to Volleyball Dataset, thus we compare our proposed PC-TDM with all of these state-of-the-art methods for group activity recognition. The comparison results are shown in Table 1. We can see that the proposed PC-TDM achieves better performance than these methods. The PC-TDM improves 5.8% and 5.2% compared with the most related work [17] on MPA and MPCA, respectively. And the PC-TDM improves 4.4% and 4.5% compared with the existing state-of-the-art performance in [30] on MPA and MPCA, respectively. The confusion matrix of PC-TDM is also shown in Figure 5. Compared to the confusion matrices reported in [17, 25], the proposed PC-TDM overcomes the confusion between class "Setting" and "Passing" well. From this experiment, we validate the effectiveness of key actors with long motion or flash motion for recognizing group activity.

Table 2: Comparison of different methods on Collective Activity Dataset. Similar to [39], the results of class "Walking" and "Crossing" are merged as "Moving".

Methods	Moving	Waiting	Queuing	Talking	MPCA
Lan *et al.* [24]	92	69	76	99	84
Choi *et al.* [6]	90	**82.9**	95.4	94.9	90.8
Zhou *et al.* [43]	88.5	74.0	95.0	98.0	88.9
Ibrahim *et al.* [17]	95.9	66.4	96.8	**99.5**	89.7
Hajimirsadeghi *et al.* [13]	87	75	92	99	88.3
Wang *et al.* [39]	94.9	63.6	**100**	**99.5**	89.4
Li *et al.* [25]	90.8	81.4	99.2	84.6	89.0
B1	97.6	51.8	**100**	**99.5**	87.2
B2	**99.5**	48.2	**100**	**99.5**	86.8
B3	92.3	73.1	**100**	**99.5**	91.2
B4	88.0	79.4	**100**	**99.5**	91.7
PC-TDM	92.8	76.6	**100**	**99.5**	**92.2**

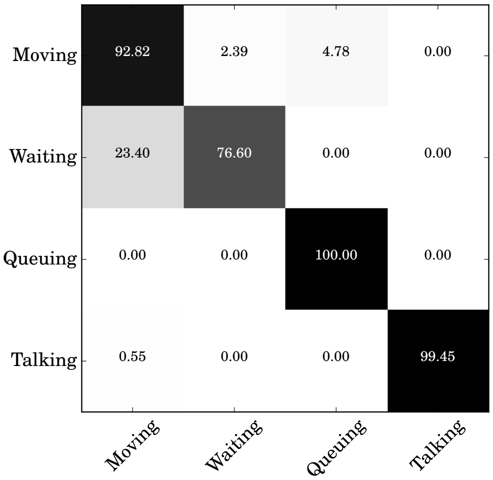
4.5 Results on Collective Activity Dataset

Comparison with baselines. Table 2 shows the recognition accuracy of the proposed PC-TDM compared with the baselines. As shown in this table, the proposed PC-TDM obtains the best performance over all baselines on MCA and MPCA respectively, due to the contributions of long motions and flash motions. Moreover, B1 and B2 also validate the effectiveness of deep feature for group activity recognition. The confusion matrix of the PC-TDM is shown in Figure 6. It is noted that the "Waiting" is confused by the "Moving" seriously, since the action class "Waiting" always occurs with class "Moving".

Comparison with state-of-the-art methods. The comparison methods include Lan *et al.* [24], Choi *et al.* [6], Zhou *et al.* [43], Ibrahim *et al.* [17], Hajimirsadeghi *et al.* [13], Wang *et al.* [39], Li *et al.* [25]. The comparisons results are shown in Table 2. It is noted that the results of these methods are calculated from the corresponding original confusion matrix in [6, 13, 17, 24, 25, 39, 43]. We can see that the proposed PC-TDM achieves the best performance, and its MPCA improves 1.4% compared with existing state-of-the-art performance in [6]. As new exploration by focusing on actions of key participants, PC-TDM improves 2.5% compared with the most related work [17] on MPCA. For "Moving", the long and flash motions are not very obvious, thus B2 without considering long and flash motions achieves the highest accuracy. Since the provided persons tracklets are not accurate, the method in [6] considering tracking achieves the highest accuracy for "Waiting". It is worth noting that B2 gets the highest accuracy on "Moving", but has the worst performance on "Waiting". To sum up, our model obtains satisfactory results on all activities, especially "Queuing" and "Talking". Thus, the MPCA of our approach is higher than the others.

5 CONCLUSIONS

In this work, we proposed a novel Participation-Contributed Temporal Dynamic Model (PC-TDM) for group activity recognition with attending to key actors (participants). The proposed PC-TDM mainly consists of an "One" Network and an "One to Key" Network. First, we utilize "One" Network to model the individual dynamics of each person from the CNN features. Second, a new "One to Key" Network (OKN) is built to progressively aggregate the motion information of key actors with long motion or flash motion over time. Specifically, OKN feeds the individual dynamics of each person into an Interaction Bi-LSTM for modeling the interaction-related dynamics according to the order of moving time of individual's long motion. Then an Aggregation LSTM is designed to aggregate the latent output states from Interaction Bi-LSTM with trainable time-varying attention weights one-by-one. In experiments, the proposed PC-TDM improves group activity recognition performance on two benchmarks (i.e., Volleyball Dataset and Collective Activity Dataset) compared with the state-of-the-art methods.



	Moving	Waiting	Queuing	Talking
Moving	92.82	2.39	4.78	0.00
Waiting	23.40	76.60	0.00	0.00
Queuing	0.00	0.00	100.00	0.00
Talking	0.55	0.00	0.00	99.45

Figure 6: Confusion matrix of the proposed PC-TDM on Collective Activity Dataset (CAD).

6 ACKNOWLEDGMENTS

This work was partially supported by the 973 Program (Grant No. 2014CB347600), the National Natural Science Foundation of China (Grant No. 61522203, 61772275, 61702265, and 61732007), the Natural Science Foundation of Jiangsu Province (Grant No. BK20170856 and BK20170033), and the CCF-Tencent Open Research Fund (PI: Xiangbo Shu).

REFERENCES

[1] Mohamed Rabie Amer, Peng Lei, and Sinisa Todorovic. 2014. HiRF: Hierarchical Random Field for Collective Activity Recognition in Videos. In *ECCV*.

[2] Jain Ashesh, Zamir Amir Roshan, Savarese Silvio, and Saxena Ashutosh. 2016. Structural-RNN: Deep Learning on Spatio-Temporal Graphs. In *CVPR*.

[3] Timur M. Bagautdinov, Alexandre Alahi, FranÃğois Fleuret, Pascal Fua, and Silvio Savarese. 2017. Social Scene Understanding: End-to-End Multi-person Action Localization and Collective Activity Recognition. In *CVPR*.

[4] Sovan Biswas and Juergen Gall. 2018. Structural Recurrent Neural Network (SRNN) for Group Activity Analysis. *arXiv preprint arXiv:1802.02091* (2018).

[5] Thomas Brox, AndrÃľs Bruhn, Nils Papenberg, and Joachim Weickert. 2004. High Accuracy Optical Flow Estimation Based on a Theory for Warping. In *ECCV*.

[6] Wongun Choi and Silvio Savarese. 2012. A unified framework for multi-target tracking and collective activity recognition. In *ECCV*.

[7] Wongun Choi, Khuram Shahid, and Silvio Savarese. 2009. What are they doing? : Collective activity classification using spatio-temporal relationship among people. In *ICCV Workshops*.

[8] Wongun Choi, Khuram Shahid, and Silvio Savarese. 2011. Learning context for collective activity recognition. In *CVPR*.

[9] Martin Danelljan, Gustav HÃďger, Fahad Khan, and Michael Felsberg. 2014. Accurate Scale Estimation for Robust Visual Tracking. In *BMVC*.

[10] Zhiwei Deng, Arash Vahdat, Hexiang Hu, and Greg Mori. 2016. Structure Inference Machines: Recurrent Neural Networks for Analyzing Relations in Group Activity Recognition. In *CVPR*.

[11] Zhiwei Deng, Mengyao Zhai, Lei Chen, Yuhao Liu, Srikanth Muralidharan, Mehrsan Javan Roshtkhari, and Greg Mori. 2015. Deep structured models for group activity recognition. In *BMVC*.

[12] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*.

[13] Hossein Hajimirsadeghi, Wang Yan, Arash Vahdat, and Greg Mori. 2015. Visual recognition by counting instances: A multi-instance cardinality potential kernel. In *CVPR*.

[14] Jun-Yan He, Xiao Wu, Yu-Gang Jiang, Bo Zhao, and Qiang Peng. 2017. Sketch Recognition with Deep Visual-Sequential Fusion Model. In *ACM Multimedia*.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

[16] Hexiang Hu, Guang-Tong Zhou, Zhiwei Deng, Zicheng Liao, and Greg Mori. 2016. Learning Structured Inference Neural Networks with Label Relations. In *CVPR*.

[17] Moustafa Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. 2016. A Hierarchical Deep Temporal Model for Group Activity Recognition. In *CVPR*.

[18] Yu-Gang Jiang, Zuxuan Wu, Jun Wang, Xiangyang Xue, and Shih-Fu Chang. 2018. Exploiting Feature and Class Relationships in Video Categorization with Regularized Deep Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 2 (2018), 352–364.

[19] Davis E. King. 2009. *Dlib-ml: A Machine Learning Toolkit.* JMLR.org. 1755–1758 pages.

[20] Diederik Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *NIPS*.

[22] Tian Lan, Lei Chen, Zhiwei Deng, Guang-Tong Zhou, and Greg Mori. 2014. Learning Action Primitives for Multi-level Video Event Understanding. In *ECCV*.

[23] Tian Lan, Leonid Sigal, and Greg Mori. 2012. Social roles in hierarchical models for human activity recognition. In *CVPR*.

[24] Tian Lan, Yang. Wang, Weilong. Yang, Stephen N. Robinovitch, and Greg Mori. 2012. Discriminative latent models for recognizing contextual group activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 8 (2012), 1549–1562.

[25] Xin Li and Mooi Choo Chuah. 2017. SBGAR: Semantics Based Group Activity Recognition. In *ICCV*.

[26] Zechao Li, Jinhui Tang, and Tao Mei. 2018. Deep Collaborative Embedding for Social Image Understanding. *IEEE transactions on pattern analysis and machine intelligence* (2018).

[27] Bingbing Ni, Shuicheng Yan, and Ashraf A. Kassim. 2009. Recognizing human group activities with localized causalities. In *CVPR*.

[28] Vignesh Ramanathan, Jonathan Huang, Sami Abuelhaija, Alexander Gorban, Kevin Murphy, and Fei Fei Li. 2016. Detecting Events and Key Actors in Multi-person Videos. In *CVPR*.

[29] Zhao Rui-Wei, Wu Zuxuan, Li Jianguo, and Jiang Yu-Gang. 2017. Learning Semantic Feature Map for Visual Content Recognition. In *ACM Multimedia*.

[30] Tianmin Shu, Sinisa Todorovic, and Song-Chun Zhu. 2017. CERN: confidence-energy recurrent network for group activity recognition. In *CVPR*.

[31] Xiangbo Shu, Jinhui Tang, Zechao Li, Hanjiang Lai, Liyan Zhang, and Shuicheng Yan. 2018. Personalized Age Progression with Bi-Level Aging Dictionary Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 4 (2018), 905–917.

[32] Xiangbo Shu, Jinhui Tang, Guo Jun Qi, Yan Song, Zechao Li, and Liyan Zhang. 2017. Concurrence-Aware Long Short-Term Sub-Memories for Person-Person Action Recognition. In *CVPR Workshops*.

[33] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *NIPS*.

[34] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

[35] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *CVPR*.

[36] Jinhui Tang, Lu Jin, Zechao Li, and Shenghua Gao. 2015. RGB-D Object Recognition via Incorporating Latent Data Structure and Prior Knowledge. *IEEE Transactions on Multimedia* 17, 11 (2015), 1899–1908.

[37] Jinhui Tang, Xiangbo Shu, Zechao Li, Guo-Jun Qi, and Jingdong Wang. 2016. Generalized deep transfer networks for knowledge propagation in heterogeneous domains. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 12, 4s (2016), 68.

[38] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, and Cordelia Schmid. 2009. Evaluation of local spatio-temporal features for action recognition. In *BMVC*.

[39] Minsi Wang, Bingbing Ni, and Xiaokang Yang. 2017. Recurrent Modeling of Interaction Context for Collective Activity Recognition. In *CVPR*.

[40] Daniel Weinland, Remi Ronfard, and Edmond Boyer. 2011. *A survey of vision-based methods for action representation, segmentation and recognition.* Elsevier Science Inc. 224–241 pages.

[41] Liang Xiaodan, Lin Liang, and Cao Liangliang. 2013. Learning latent spatio-temporal compositional model for human action recognition. In *ACM Multimedia*.

[42] Yichao Yan, Bingbing Ni, and Xiaokang Yang. 2017. Predicting Human Interaction via Relative Attention Model. In *IJCAI*.

[43] Zheng Zhou, Kan Li, Xiangjian He, and Mengmeng Li. 2016. A generative model for recognizing mixed group activities in still images. In *IJCAI*.