Coherence Constrained Graph LSTM for Group Activity Recognition

Jinhui Tang, Xiangbo Shu, Rui Yan, and Liyan Zhang

Abstract—This work aims to address the group activity recognition problem by exploring human motion characteristics. Traditional methods hold that the motions of all persons contribute equally to the group activity, which suppresses the contributions of some relevant motions to the whole activity while overstating some irrelevant motions. To address this problem, we present a Spatio-Temporal Context Coherence (STCC) constraint and a Global Context Coherence (GCC) constraint to capture the relevant motions and quantify their contributions to the group activity, respectively. Based on this, we propose a novel Coherence Constrained Graph LSTM (CCG-LSTM) with STCC and GCC to effectively recognize group activity, by modeling the relevant motions of individuals while suppressing the irrelevant motions. Specifically, to capture the relevant motions, we build the CCG-LSTM with a temporal confidence gate and a spatial confidence gate to control the memory state updating in terms of the temporally previous state and the spatially neighboring states, respectively. In addition, an attention mechanism is employed to quantify the contribution of a certain motion by measuring the consistency between itself and the whole activity at each time step. Finally, we conduct experiments on two widely-used datasets to illustrate the effectiveness of the proposed CCG-LSTM compared with the state-of-the-art methods.

Index Terms—group activity recognition, long short-term memory, fine-grained motion, deep learning.

1 Introduction

TRADITIONAL action recognition, such as single-person action recognition [1] and two persons' interaction recognition [2], usually performed by one/two persons in a video, has achieved satisfactory performance over the past decades [3], [4]. Compared with traditional human action, group activity [5] (also called collective activity [6]) is a more complex yet common action in a scene. Different from single-person action and two persons' interaction, a group activity is usually performed by multiple ($\geq$ 3) persons simultaneously. Therefore, in group activity recognition, we need to model multiple persons' individual actions and their interactions. This is a fine-grained recognition task compared to traditional single-person action recognition and two persons' interaction recognition [7], and thus to be much more challenging.

Benefit from the success of Recurrent Neural Network (RNN) [8], especially for Long Short Term Memory (LSTM) [9], group activity recognition has made progress in some extent in recent years [10], [11], [12], [13]. By reviewing existing deep learning methods related to group activity recognition, a common solution is to first learn a person-level action representation of each person, and then integrate all the individual representations to recognize the group-level activity. Specifically, some early methods assumed that all persons in an activity scene are independent from each other [10], [11]. Subsequently, some works consider that all persons in an activity scene are dependent on each other, and model each person's individual motion by referring to the other persons' motion states [12], [2], [14].

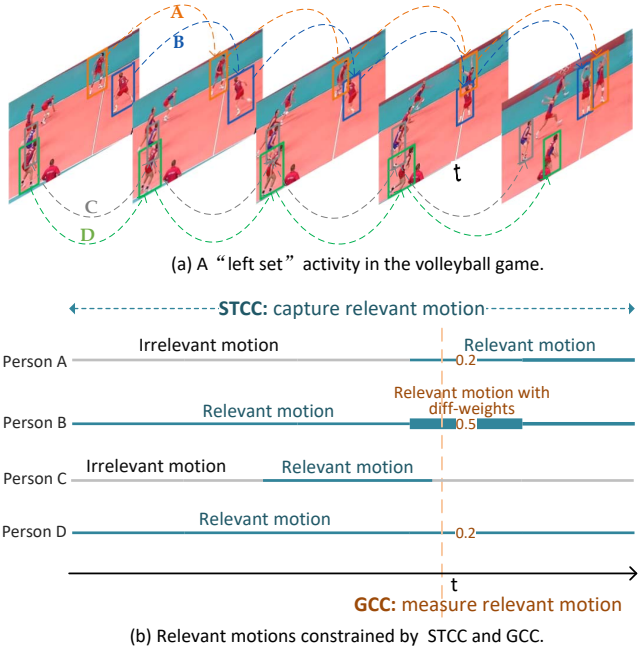


Fig. 1. Illustration of the relevant motions constrained by Spatio-Temporal Context Coherence (STCC) and Global Context Coherence (GCC) in a "Left set" activity of volleyball game. STCC aims to capture the relevant motions of persons, and GCC aims to measure the contribution of the relevant motions.

However, the aforementioned methods hold that the motions of all persons contribute equally to the group activity, which suppresses the contributions of some relevant motions to the whole activity and overstates some outlier

J. Tang, X. Shu, and R. Yan are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China. E-mail:{jinhuitang, shuxb, ruiyan}@njust.edu.cn. (Corresponding author: Xiangbo Shu)
L. Zhang is with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China. E-mail: zhangliyan@nuaa.edu.cn.

motions irrelevant to the whole activity. To address this issue, Deng et al. [13], [15] indicated that most persons contribute to the group activity, while some persons are irrelevant to the group activity. They explored all the motion information of "relevant" persons to recognize the group activity in an iterative manner while ignoring all the motion information from "irrelevant" persons. Unfortunately, the "relevant" persons are not always relevant to the group activity, while the "irrelevant" persons are not always irrelevant to the group activity. That is to say, the contribution of a certain motion to the whole activity is decided only by its relevance to the activity, regardless of the person it comes from.

Therefore, how to capture the relevant motions of individuals becomes important for understanding group activity. By observing the group activities, we find that: 1) in the temporal domain, most of the motions of a certain person are usually coherent most of the time; 2) in the spatial domain, the motion of a certain person is usually consistent with the context motion information from the other persons most of the time. Based on these observations, we present a **Spatio-Temporal Context Coherence (STCC)** constraint: if an individual's motion is coherent in the temporal domain and consistent with other individuals' motions in the spatial domain, this motion belongs to the relevant motion. For example, in Figure 1, most of the relevant motions of individuals are relevant to the "left set" activity. These relevant motions are not only consistent with each other in the spatial domain, but also coherent to their previous motions in the temporal domain. Obviously, the spatial interactions among persons are dependent on the temporal motions of individuals. Recently, it has been proven that Graph LSTM can model sequence data in spatio-temporal domains [16] simultaneously. Motivated by this, we consider extending Graph LSTM with the STCC constraint to understand the group activity by exploring the individual motions in both spatial and temporal domains.

Empirically, although all relevant motions of individuals contribute to inferring the class of group activity, their contributions are different. To better illustrate this, we take the "left set" activity in a volleyball game as an example, as shown in Figure 1. Both the "setting" motion of person B and the "moving" motion of person D are relevant motions in the "left set" activity. However, the former motion is more crucial than the latter motion. Thus, to better understand the group activity, we should learn to quantitatively measure the contribution of a certain motion to the whole activity at a certain time step. To this end, we present another critical constraint, namely **Global Context Coherence (GCC)**: the more a certain motion is consistent with the whole activity, the larger contribution it makes, and vice versa. Inspired by the attention models in previous works [17], [18], we adopt an attention mechanism to quantify the contribution of a certain motion by measuring the consistency between itself and the whole activity under the GCC constraint.

To explore the relevant motions while suppressing the irrelevant motions, we propose a novel **Coherence Constrained Graph LSTM (CCG-LSTM)** with STCC and GCC constraints for group activity recognition. It is shown in Figure 2 by using a volleyball game as an example. First, we extract the CNN feature of each person on the detected and tracked bounding box by employing a pre-trained CNN model [19]. Second, we take the CNN features of all persons as the input of CCG-LSTM, to jointly learn the individual motion states of all persons under the STCC constraint over time. Specifically, to capture the relevant motions at each time, a CCG-LSTM Unit with a spatial confidence gate and a temporal confidence gate is built to control the memory state updating in terms of the temporally previous motion state and the spatially neighboring motion states. Third, we employ an attention mechanism with GCC to quantify the contribution of the relevant motions by learning different attention factors corresponding to different motions. Here, a specific attention factor of a certain motion measures the contribution of this motion to the whole activity at a certain time step. Subsequently, at each time step, an Aggregation LSTM aggregates all the individual motion states weighted by different attention factors into a hidden representation of the whole activity. After that, each hidden representation of the activity is input to the softmax classifier at each time step, and then we average the outputs of all the softmax classifiers to infer the class of the group activity.

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

- We deeply explore the human motion characteristics in the scenario of group activity with multiple persons, and present two constraints, i.e., Spatio-Temporal Context Coherence (STCC) and Global Context Coherence (GCC), to capture and quantify the relevant motions of individuals.
- To effectively recognize group activities, we propose a novel Coherence Constrained Graph LSTM (CCG-LSTM) to learn the discriminative representation of a whole activity by modeling the motions of individuals relevant to the whole activity, while suppressing the irrelevant motions.
- We conduct experiments on two widely-used datasets (Volleyball Dataset [10] and Collective Activity Dataset [20]) to illustrate the effectiveness of the proposed CCG-LSTM method compared with the state-of-the-art methods.

The rest of this paper is organized as follows. Section 2 reviews related works. Section 3 introduces the proposed CCG-LSTM in details. Experiments are conducted in Section 4, followed by the conclusion in Section 5.

2 Related Works

2.1 RNN-based Action Recognition

Action recognition aims to recognize an activity performed by one/two persons in the computer vision filed [21], [22], [2], [23]. A large family of action recognition methods provided various spatio-temporal features to represent the action in a video, such as Histogram of Oriented Gradients (HOG) [24], Dense Trajectories [22], 3D-SIFT [25], Histogram of Optical Flow (HOF) [26], and so on.

As a neural network for handling sequential data with variable length [27], [28], Recurrent Neural Networks (RNN) [8], especially for Long Short-Term Memory (LSTM) [9], have made progress in action recognition for the last five years [29], [21], [30], [31], [11], [10]. In the early stage,

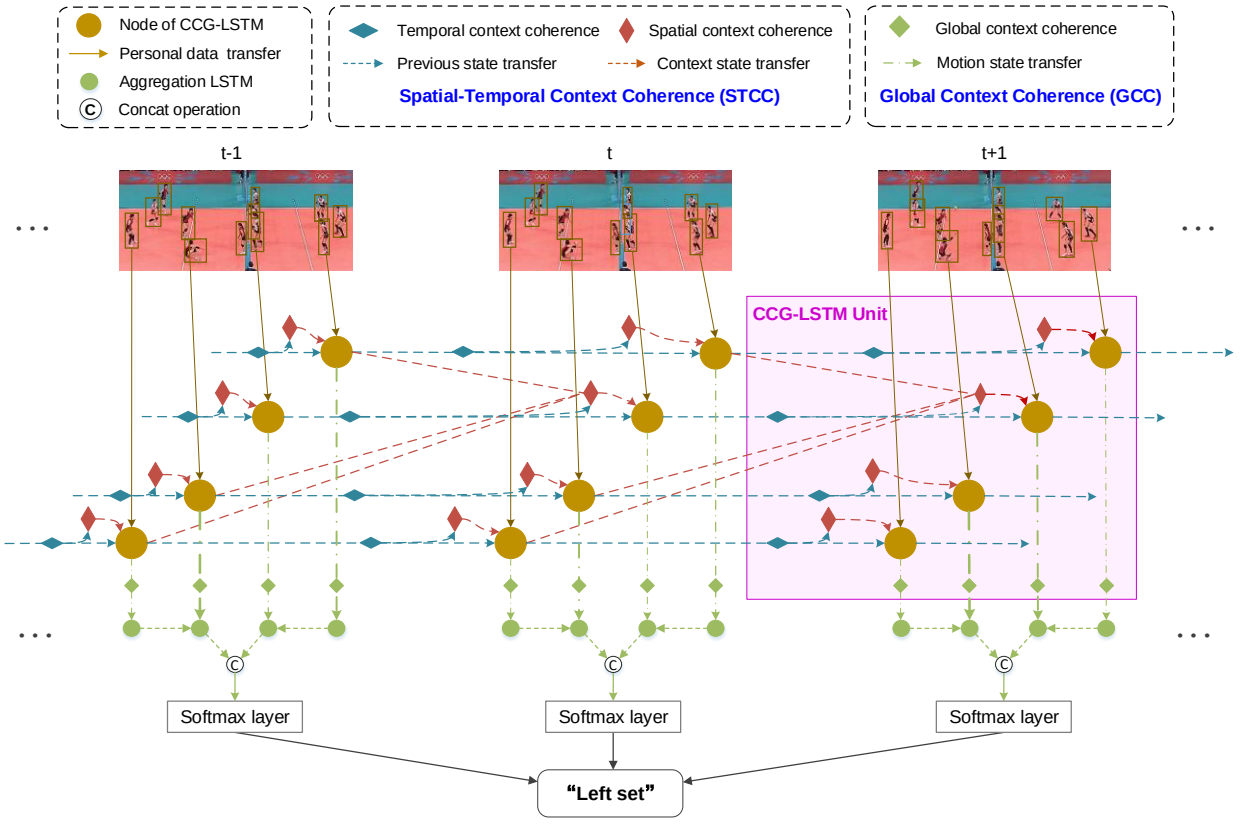


Fig. 2. The proposed Coherence Constrained Graph LSTM (CCG-LSTM) for recognizing a group activity. The personal data transfer, previous state transfer, context state transfer and motion state transfer denote the transferring of the input feature of individual, previous motion state of individual, spatial context state of neighbor, and current motion sate of individual, respectively. Spatial-Temporal Context Coherence (STCC) consists of the spatial context coherence and spatial context coherence. Aggregation LSTM aggregates the relevant motions of all persons with different attention factors into the hidden representation of the whole activity at each time step. Best view in color version.

some researchers employed an RNN as a temporal-sequence classifier to learn the representation of action by capturing the temporal information with frames over time [21], [32]. The common solution is to combine the CNN layers and RNN/LSTM layers in a bottom-up way. For example, Wu *et al.* [33] proposed to train three types of CNNs equipped with an LSTM layer to model the spatial, short-term motion and audio corresponding to the inputs of video frames, stacked optical flows, and audio spectrograms, respectively.

To leverage the spatial information between different body parts, Du *et al.* [30] divided a human body skeleton into five parts based on the human skeleton, and then fed these five parts into five RNNs. As the number of layers increases, the representation outputs from multiple RNNs are hierarchically fused into the inputs of the higher layers. Subsequently, for the spatial interaction information among different persons, Shu *et al.* [2] proposed to model the long-term inter-related motions among interacting individuals, rather than the individual motions of each person.

Moreover, for various action scenarios [29], [34], some researchers evolved the architectures of the traditional LSTM to address the problem of action recognition well. For example, to capture the temporal change of motion information between two consecutive frames, Veeriah *et al.* [31] proposed a Differential RNN architecture equipped with the Derivative of States between LSTM gates. Moreover, Shahroudy *et al.* [34] proposed a Part-aware LSTM that separates the memory cell into multiple sub-cells corresponding to different skeleton parts and explicitly models the dependencies over spatial and temporal domains concurrently.

2.2 Group Activity Recognition

In contrast to traditional action recognition, group activity recognition aims to automatically understand an activity performed by at least three persons [20], [35], [36], [37]. Over the past years, group activity recognition has developed into an attractive topic in the computer vision area [6], [15], [38], [39], [40], [41]. Since Deep Neural Networks (DNN) have shown excellent performance in a variety of computer vision tasks, many DNN-based activity recognition methods have been proposed in recent years [10], [15], [42], [43]. As one of most representative works, Ibrahim *et al.* [10] proposed a hierarchical model with several LSTM layers to learn the motion state of each person in a temporal sequence, and then combine the motion states of all persons into the hidden presentation of the whole activity in each frame. Similarly, Wang *et al.* [42] extended an RNN-based hierarchical framework to learn three level motions, i.e., person-level motions, group-level motions and scene-level motions corresponding to the individuals, persons within a group, persons within at least two groups.

In a group activity, most persons interact with each other, which is the main characteristic compared with the single-person action. Therefore, some works [44], [7]

proposed to model the interaction-related motions among persons over time. For example, Shu *et al.* [44] proposed a Confidence-Energy Recurrent Network to integrate the related-confidences from two types of predictions (individual action prediction and human interaction prediction) into an energy layer in inferring the class of event, i.e., event detection [45].

However, existing methods hold that all persons in a group activity contribute equally to this activity, which brings in some irrelevant motions of individuals to the whole activity. Recently, Deng *et al.* [15], [13] indicated that most of the persons contribute to inferring the class of the group activity, while a small number of persons are irrelevant to the activity most of the time. Specifically, Deng *et al.* [15], [13] proposed an inference learning model to iteratively find the "relevant" persons, while removing the "irrelevant" persons based on the relation between the person-level class label and group-level class label. Unfortunately, the "relevant" persons are not always relevant to the group activity, while the "irrelevant" persons are not always irrelevant to the group activity. Therefore, it is more reasonable to capture the relevant motions themselves, no matter they come from the "relevant" persons or "irrelevant" persons. To explore the relevant motions while suppress the irrelevant motions, we propose a novel Coherence Constrained Graph LSTM (CCG-LSTM) with STCC and GCC constraints.

3 Coherence Constrained Graph LSTM

3.1 Motivation

For traditional action recognition, given a video clip $\{\mathbf{x}^t \in \mathbb{R}^D | t = 1, \cdots, T\}$ with T frames, where $\mathbf{x}^t$ is the static feature (such as a CNN feature [19]) of the t-th frame, and D is the dimension of $\mathbf{x}^t$, we can utilize Long Short-Term Memory (LSTM) [9] to learn a sequence of motion states $\{\mathbf{h}^t \in \mathbb{R}^d | t = 1, \cdots, T\}$ to describe a person's action or multiple persons' activity in this video clip. For a group activity [10], multiple persons ($\geq$ 3 persons) interact with each other in the spatial domain, and their motions vary with time in the temporal domain. We analyze the motions of all persons from time step 1 to T, and recognize what they are doing in this video clip, which is called group activity recognition in this paper. If we directly employ the traditional LSTM [9] to learn the representation of the whole activity based on the frame-level features, some specific motion information of individuals cannot be captured well, and the learned representations of different activities are not discriminative enough. Thus, similar to the recent LSTM-based methods [10], [44], we model the group activity via the person-level features rather than the frame-level features.

It has been proven that LSTM can well capture the temporal motions of individuals [10], [21], while some extended approaches were proposed to model the spatial interactions among individuals [32], [2]. However, they neglect the fact that the spatial interactions among persons are dependent on the temporal motions of individuals, which is an important clue for group activity recognition. As we know, the graph-based learning methods can build the spatial relation of a certain number of nodes [46], [47]. Some researchers integrate the ideas of graph-based learning and LSTM to propose Graph LSTM methods for object detection [48] and image segmentation [49], by regarding the super-pixels of image as a sequential set in the spatial domain. This kind of method models each sequential data via an LSTM, where each LSTM unit is regarded as a graph node, and each edge connects two nodes. Motivated by this, we make an attempt to model the temporal motions of individuals and spatial interactions among persons concurrently via the Graph LSTM in this work.

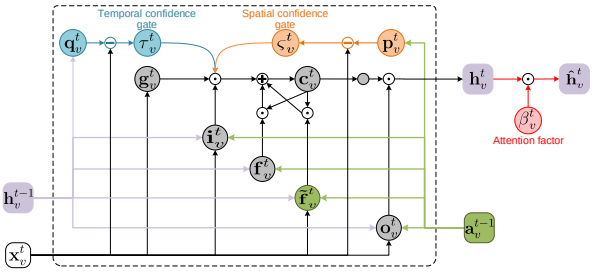


Fig. 3. A node of the CCG-LSTM at time step t. The components marked with blue color, orange color and red color denote the Temporal Context Coherence (TCC), Spatial Context Coherence (SCC), and Global Context Coherence (GCC) constraints, respectively.

3.2 Graph LSTM

Given a video clip with T frames describing a specific group activity within V persons, let $\mathbf{x}_v^t \in \mathbb{R}^D$ denote the static feature (e.g., CNN features) of the v-th person in the t-th frame, where $t \in \{1, 2, \cdots, T\}$ and $v \in \{1, 2, \cdots, V\}$, we can construct a spatial sequential set $\{\mathbf{x}_v^t\}_{v=1}^V$ for each frame and a temporal sequential set $\{\mathbf{x}_v^t\}_{t=1}^T$ for each person. These two types of sets in the spatial and temporal domains can be used to construct a set of graphs $\mathcal{G}^t = \{\mathcal{S}^t, \mathbf{E}^t\}$ ($t = 1, 2, \cdots, T$), where $\mathcal{S}^t = \{\mathbf{x}_v^t\}_{v=1}^V$ is the set of nodes at time step t, $\mathbf{E}^t$ is the adjacency-edge matrix, $\mathbf{x}_v^t$ is the v-th node of $\mathcal{G}^t$ corresponding to the v-th person, and $\mathbf{E}_{i,j}^t$ denotes an adjacency edge between the i-th node and the j-th node in graph $\mathcal{G}^t$. Similar to [50], we extend the traditional LSTM to a Graph LSTM. In the Graph LSTM, for the v-th node, its input gate $\mathbf{i}_v^t$, forget gate $\mathbf{f}_v^t$, output gate $\mathbf{o}_v^t$ and its neighboring forget gate $\tilde{\mathbf{f}}_v^t$ at time step t are decided by its input feature $\mathbf{x}_v^t$ at current time step, its motion state $\mathbf{h}_v^{t-1}$ and the spatial context state $\mathbf{a}_v^{t-1}$ from its spatial neighbors at the previous time step, respectively. Formally, at time step t, the motion state $\mathbf{h}_v^t$ of the v-th node in Graph LSTM can be formulated as follows.

$$\mathbf{a}_v^{t-1} = \left(\mathbf{E}_{:,v}^t\right)^{\mathrm{T}} [\cdots, \mathbf{h}_i^{t-1^{\mathrm{T}}}, \cdots]^{\mathrm{T}} + \mathbf{b}_a, \ i \in \Phi(v); \quad (1)$$

$$\mathbf{i}_v^t = \sigma(\mathbf{W}_i \mathbf{x}_v^t + \mathbf{U}_i \mathbf{h}_v^{t-1} + \mathbf{G}_i \mathbf{a}_v^{t-1} + \mathbf{b}_i); \quad (2)$$

$$\mathbf{f}_v^t = \sigma(\mathbf{W}_f \mathbf{x}_v^t + \mathbf{U}_f \mathbf{h}_v^{t-1} + \mathbf{G}_f \mathbf{a}_v^{t-1} + \mathbf{b}_f); \quad (3)$$

$$\tilde{\mathbf{f}}_v^t = \sigma(\mathbf{W}_{\tilde{f}} \mathbf{x}_v^t + \mathbf{U}_{\tilde{f}} \mathbf{h}_v^{t-1} + \mathbf{G}_{\tilde{f}} \mathbf{a}_v^{t-1} + \mathbf{b}_{\tilde{f}}); \quad (4)$$

$$\mathbf{o}_v^t = \sigma(\mathbf{W}_o \mathbf{x}_v^t + \mathbf{U}_o \mathbf{h}_v^{t-1} + \mathbf{G}_o \mathbf{a}_v^{t-1} + \mathbf{b}_o); \quad (5)$$

$$\mathbf{g}_v^t = \varphi(\mathbf{W}_g \mathbf{x}_v^t + \mathbf{U}_g \mathbf{h}_v^{t-1} + \mathbf{G}_g \mathbf{a}_v^{t-1} + \mathbf{b}_g); \quad (6)$$

$$\mathbf{c}_s^{t-1} = \frac{\sum_{i \in \Phi(v)} \mathbf{c}_i^{t-1}}{|\Phi(v)|}; \quad (7)$$

$$\mathbf{c}_v^t = \mathbf{i}_v^t \odot \mathbf{g}_v^t + \mathbf{f}_v^t \odot \mathbf{c}_v^{t-1} + \tilde{\mathbf{f}}_v^t \odot \mathbf{c}_s^{t-1}; \quad (8)$$

$$\mathbf{h}_v^t = \mathbf{o}_v^t \odot \varphi(\mathbf{c}_v^t), \quad (9)$$

where $\mathbf{W}_*$, $\mathbf{U}_*$, and $\mathbf{G}_*$ are weight matrices, $\mathbf{b}_*$ is a bias vector, $\sigma(\cdot)$ is the sigmoid function, $\varphi(\cdot)$ is the hyperbolic tangent $tanh(\cdot)$, $\odot$ denotes element-wise product, $\Phi(v)$ denotes the set of the neighbors of the v-th node in graph $\mathcal{G}^t$, and $\mathbf{c}_s^{t-1}$ denotes the spatial context memory state of the v-th node, which is the average of the neighboring memory states.

As we discussed before, most of the motions are relevant to the whole activity, while a small number of motions are irrelevant in a group activity. For the relevant motions, their contributions to the group activity are different. Thus, the main goal of this work is to find the crucial relevant motions and measure their contributions for inferring the class of group activity. Different from traditional group activity recognition methods treating the motions equally, the motions in this work are regarded as fine-grained motions since they have different contributions to the group activity. In this work, we present a Spatio-Temporal Context Coherence (STCC) constraint and a Global Context Coherence (GCC) constraint to capture and quantify the relevant motions of individuals, respectively. By extending the basic Graph LSTM, we propose a Coherence Constrained Graph LSTM (CCG-LSTM) with STCC and GCC constraints.

The details of a certain node of CCG-LSTM at time step t are shown in Figure 3. The components marked with blue color, orange color and red color denote the Temporal Context Coherence (TCC), Spatial Context Coherence (SCC), and Global Context Coherence (GCC) constraints, respectively. The details of the TCC, SCC, and GCC are introduced in the following sections.

3.3 Spatio-Temporal Context Coherence

As aforementioned, at a certain time step, if one person's motion is coherent to her/his motions in the temporal domain, as well as consistent with other persons' motions in the spatial domain, this kind of motions is relevant motion. We call it Spatio-Temporal Context Coherence (STCC) constraint in this paper. STCC consists of Temporal Context Coherence (TCC) and Spatial Context Coherence (SCC) corresponding to the temporal and spatial domains, respectively. Here, we first introduce SCC in detail.

Generally, a group activity contains multiple persons who perform their respective actions over time. The primary consideration of the group activity recognition problem is how to learn the individual motion state of each person. The common strategy in recent methods [2], [10], [42], [44] employs the sequence models (e.g., RNN and LSTM) to model the individual motions from T consecutive frames, which is in accordance with the kinematic principle. To better understand the individual motion, we bring in a new characteristic of individual motion based on **Temporal Context Coherence (TCC)**: most of the relevant motions of a certain person are coherent over time.

Intuitively, TCC reflects that the relevant motions of a certain person between two consecutive frames are similar to each other. Thus, we can employ the TCC constraint to exclude the sudden motions to some extent. For example, in the "left set" activity of a volleyball game, a person who is moving over most of the time steps suddenly falls down. This "fall down" motion is not coherent to the previous motions. It should be suppressed when updating the memory state in CCG-LSTM.

Formally, to bring in the TCC constraint into CCG-LSTM, we design a temporal confidence gate (denoted by τ_v^t) to control the motion information transferring across the CCG-LSTM units (CCG-LSTM unit is an unfolded part of CCG-LSTM at one time step) in the temporal domain. To measure the motion coherence of a certain person at a certain time step, we adopt the similarity between the previous motion state and the current input feature.

Since the dimensions of the input feature and the motion state are different, we need to project them into a common space. For the v-th person at time step t, the projection vector of input feature $\mathbf{x}_v^t$ can be computed by

$$\tilde{\mathbf{x}}_v^t = \varphi(\mathbf{W}_x(\mathbf{x}_v^t)), \quad (10)$$

where $\mathbf{W}_x : \mathbb{R}^D \rightarrow \mathbb{R}^M$ is a projection matrix. Similarly, the projection vector of the motion state $\mathbf{h}_v^{t-1}$ of the v-th person at the previous time step can be computed as,

$$\mathbf{q}_v^t = \varphi(\mathbf{W}_q(\mathbf{h}_v^{t-1})), \quad (11)$$

where $\mathbf{W}_q : \mathbb{R}^d \rightarrow \mathbb{R}^M$ is a projection matrix.

The temporal confidence gate τ_v^t at time step t is activated by the difference between $\tilde{\mathbf{x}}_v^t$ and $\mathbf{q}_v^t$, as follows,

$$\boldsymbol{\tau}_v^t = \frac{1}{\exp(\rho(\tilde{\mathbf{x}}_v^t - \mathbf{q}_v^t)^2)}, \quad (12)$$

where the parameter ρ controls the bandwidth of the function.

Finally, under the TCC constraint, the updating equation (Eq.(8)) of memory state $\mathbf{c}_v^t$ at time step t becomes

$$\mathbf{c}_v^t = \boldsymbol{\tau}_v^t \odot \mathbf{i}_v^t \odot \mathbf{g}_v^t + \mathbf{f}_v^t \odot \mathbf{c}_v^{t-1} + \tilde{\mathbf{f}}_v^t \odot \mathbf{c}_v^{t-1}. \quad (13)$$

In the above updating equation of the memory state $\mathbf{c}_v^t$, if the current motion state is incoherent to the previous motion state, it is an irrelevant motion, which is suppressed by the temporal confidence gate $\boldsymbol{\tau}_v^t$ when updating the memory state $\mathbf{c}_v^t$.

Beyond the temporal motions of individuals, the interactions among persons should also be considered in a group activity. In other words, the relevant motions of different persons are not only coherent to their motions at the previous time steps in the temporal domain, but also consistent with each other in the spatial domain. Therefore, besides TCC, we present a **Spatial Context Coherence (SCC)** constraint to further capture the relevant motions of individuals in the spatial domain, by assuming that the relevant motions of all persons are consistent with each other in the spatial domain.

Intuitively, SCC reflects that the relevant motions of a certain person are consistent with the surrounding persons in the spatial domain. For example, in Figure 4(b), the relevant motions of the left three persons are walking together, which are consistent with each other. However, the motions of the rightmost person are irrelevant motions, which are not consistent with the motions of the other persons in the same scene. Thus, SCC is crucial for capturing the relevant motions of individuals in the spatial domain.

Formally, we design a spatial context confidence gate (denoted by ς_v^t) to bring the SCC constraint into CCG-LSTM. This gate controls the motion information transferring across CCG-LSTM in the spatial domain. In this work, the spatial confidence gate is activated by the consistency between the input feature of a certain person and the spatial context state from her/his neighbors. Specifically, for the v-th person, her/his spatial context state is defined as

$$\mathbf{a}_v^{t-1} = \sum_{i \in \Phi(v)} e_{v,i}^{t-1} \mathbf{h}_i^{t-1}, \tag{14}$$

where $e_{v,i}^{t-1}$ denotes the relationship weight between the v-th person and the i-th person at time step t. Here, it is learned via $e_{v,i}^{t-1} = \text{softmax}(\phi(h_v^{t-1}, h_i^{t-1}))$, where h_v^{t-1} denotes the v-th person's motion state at time step $(t-1)$ and $\phi(\cdot)$ is the multi-layer perceptron in [51]. Similar to Eq. (10), the projection vector of the spatial context state is computed as

$$\mathbf{p}_v^t = \varphi(\mathbf{W}_p(\mathbf{a}_v^{t-1})), \tag{15}$$

where $\mathbf{W}_p : \mathbb{R}^d \rightarrow \mathbb{R}^M$.

Based on the similarity between the input feature and the spatial context state of the v-th person, the spatial context confidence gate ς_v^t can be activated via

$$\varsigma_v^t = \frac{1}{\exp(\rho(\tilde{\mathbf{x}}_v^t - \mathbf{p}_v^t)^2)}. \tag{16}$$

Then, under the GCC constraint, the memory state $\mathbf{c}_v^t$ at time step t is updated as

$$\mathbf{c}_v^t = \varsigma_v^t \odot \mathbf{i}_v^t \odot \mathbf{g}_v^t + \mathbf{f}_v^t \odot \mathbf{c}_v^{t-1} + \tilde{\mathbf{f}}_v^t \odot \mathbf{c}_v^{t-1}. \tag{17}$$

In the above updating equation of the memory state $\mathbf{c}_v^t$, if the input feature of a certain person is inconsistent with her/his spatial context state, it is an irrelevant motion, and is suppressed during memory state updating.

When Spatio-Temporal Context Coherence (STCC) is equipped to CCG-LSTM, the motions of individuals are jointly constrained by the TCC, and SCC. In STCC, if the motion of a certain person is coherent in the temporal domain and consistent with the other individuals' motions in the spatial domain, this motion is relevant to the whole activity. Therefore, in CCG-LSTM with STCC, the memory state updating of $\mathbf{c}_v^t$ at time step t is jointly constrained by the temporal confident gate $\boldsymbol{\tau}_v^t$ and the spatial confident gate ς_v^t, as follows

$$\mathbf{c}_v^t = \boldsymbol{\tau}_v^t \odot \varsigma_v^t \odot \mathbf{i}_v^t \odot \mathbf{g}_v^t + \mathbf{f}_v^t \odot \mathbf{c}_v^{t-1} + \tilde{\mathbf{f}}_v^t \odot \mathbf{c}_v^{t-1}. \tag{18}$$

The above equation can be explained as follows. For a certain person, if her/his current input feature $\mathbf{x}_v^t$ is different from her/his previous motion state and spatial context state, the value of $\boldsymbol{\tau}_v^t \odot \varsigma_v^t$ is small, and the current motion state of this person is strongly suppressed during memory state updating.

3.4 Global Context Coherence

When we capture the relevant motion states of individuals (i.e., individual representations) under the STCC constraint, a straightforward strategy is to fuse these motion states into the hidden representation of activity via the pooling operation, e.g., max pooling, and average pooling. For example, Shu *et al.* [10] adopted the average pooling strategy

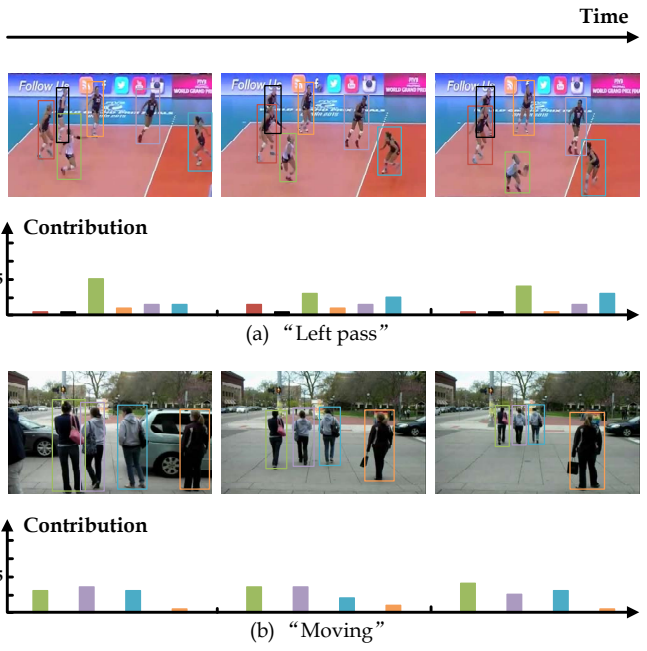


Fig. 4. Some examples of relevant motion quantified by Global Context Coherence (GCC). The higher bar chart indicates the corresponding motion is more relevant to the group activity.

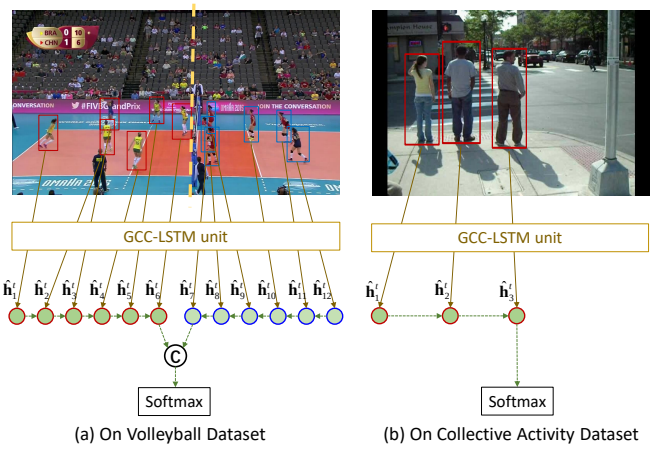


Fig. 5. Examples of Aggregation LSTM on VD and CAD. (a) and (b) denote the implementation of Aggregation LSTM in the volleyball match with two sub-groups, and the collective activity with a group, respectively.

to combine a group of individual motion states into the hidden representation of activity. We find that this strategy holds that the motions of all persons contribute equally to the group activity.

In fact, although all relevant motions of individuals contribute to inferring the class of group activity, their contributions are different. Although STCC can capture the relevant motions of individuals, it cannot quantify the contributions of the relevant motions to the group activity. Therefore, we present a **Global Context Coherence (GCC)** constraint to quantify the contributions of the relevant motions in a learning way. As aforementioned in the Introduction section, the GCC is defined as: the more a certain motion is consistent with the whole activity, the larger contribution it makes.

Therefore, the consistency between the motion of a cer-

tain person and the whole activity is the key point in GCC. Inspired by the attention models in previous works [17], [18], we adopt an attention mechanism to measure the consistency between itself and the whole activity. However, the hidden representation of the whole activity is the target we aim to learn, and thus it is unknown. Therefore, we use the average motion state $\mathbf{h}_f^t$ of all individuals' motion states to approximate the hidden representation of the whole activity in this part, namely $\mathbf{h}_f^t = \frac{1}{V}\sum_{i=1}^{V}\mathbf{h}_i^t$. Then we employ an attention model to learn an attention factor β_v^t that measures the contribution of the motion state $\mathbf{h}_v^t$ of the v-th person to the whole activity,

$$\beta_v^t = \text{softmax}(\phi(\mathbf{h}_v^t, \mathbf{h}_f^t); \gamma),\ v = 1, 2, \cdots, V, \quad (19)$$

where γ is a temperature parameter [17]. We show some examples of relevant motion quantified by GCC in Figure 4. The attention model in [17] directly use all known features as the input of the multi-layer perceptron to compute the attention factor of this person. It can be seen that the attention model of GCC does not require the known representation feature of the activity scene, in contrast to the attention model in [17].

Then, we can compute the motion state $\hat{\mathbf{h}}_v^t$ of the v-th person under the GCC constraint by

$$\hat{\mathbf{h}}_v^t = \beta_v^t \mathbf{h}_v^t. \quad (20)$$

To date, the obtained $\hat{\mathbf{h}}_v^t$ is the final motion state of the v-th person at the time step t in the Graph LSTM, under the constraints of STCC and GCC.

Similar to [36], we employ an Aggregation LSTM in the spatial domain to aggregate the motion states of all persons into a hidden representation of the whole activity at time step t person-by-person, as shown in Figure 5. We take all persons' motion states $\{\hat{\mathbf{h}}_v^t\}_{v=1}^{V}$ at time step t as the input of the Aggregation LSTM, i.e.,

$$\mathbf{z}^t = \text{Aggregation\_LSTM}(\hat{\mathbf{h}}_v^t, \mathbf{s}_{v-1}^t), \quad (21)$$

where $\mathbf{s}_{v-1}^t$ is the hidden state in Aggregation LSTM, and $\mathbf{z}^t$ is the hidden representation of the whole activity at time step t.

Finally, we feed every $\mathbf{z}^t$ ($t = 1, 2, \cdots, T$) into a softmax classifier, i.e.,

$$\mathbf{y}^t = \text{softmax}(\mathbf{z}^t),\ t = 1, 2, \cdots, T, \quad (22)$$

and then we average the outputs of all the softmax classifiers to obtain the probability class vector of group activity.

4 Experiments

We conduct experiments on two widely-used benchmarks to validate the effectiveness of the proposed CCG-LSTM compared with the state-of-the-art methods.

4.1 Baselines

For the ablation studies of the proposed CCG-LSTM, we present seven baseline methods, as follows.

- **B1** **Frame-Level CNN**. This baseline is the basic CNN model fine-tuned for group activity recognition in frames without considering the individuals. The static CNN feature of each frame is input into the softmax classifier, and we average the outputs of all softmax classifiers to infer the class of the group activity.
- **B2** **Person-Level CNN**. This baseline uses the pre-trained CNN model to extract the fc7 feature of each person on the person bounding box. Then, the features of all persons are max-pooled into a single feature at each time step. Finally, each of this single feature is input into the softmax classifier, and we average the outputs of all softmax classifiers to infer the class of the group activity.
- **B3** **Graph LSTM**. If the proposed CCG-LSTM is without any coherent constraint, it becomes a basic Graph LSTM model. Specifically, we feed the CNN features of each person into the Graph LSTM, and then pool features of all persons as the latent representation of each frame to train a softmax classifier. This baseline is designed to illustrate the capability of Graph LSTM for modeling the temporal motions and spatial interactions.
- **B4** **CCG-LSTM with only TCC**. This baseline is a simple version of the proposed CCG-LSTM while considering only the TCC, and its implementation is similar to CCG-LSTM. This baseline aims to illustrate the importance of the TCC for capturing relevant motions in the temporal domain.
- **B5** **CCG-LSTM with only SCC**. This baseline is a simple version of the proposed CCG-LSTM with only considering the SCC, and its implementation is similar to CCG-LSTM. This baseline can illustrate the importance of the SCC for capturing relevant motions in the spatial domain.
- **B6** **CCG-LSTM with only STCC**. To illustrate the effectiveness of STCC, we design this important baseline by omitting GCC in CCG-LSTM. This baseline captures the relevant motions of all persons equally.
- **B7** **CCG-LSTM with only GCC**. To illustrate the effectiveness of GCC, we designed this baseline to directly measure all motions (relevant or not) without considering the STCC and obtain the latent representation of each frame with GCC.

4.2 Implementation Details

For each video clip, we track a set of bounding boxes (tracklets) around each person over $T = 10$ time steps by the object tracker [52] in the Dlib library [53]. To address the problem of person missing in some frames, we adopt the simple strategy used in [10], [36] to make up the feature of the missing person by a full-zero matrix.

Similar to [2], [10], [44], we train the proposed CCG-LSTM in a stage-wise manner. Specifically, we train a CNN model to recognize individuals' actions, and extract the CNN features of individuals on the person's bounding boxes, which are input into CCG-LSTM for group activity recognition. The proposed CCG-LSTM is compatible with various networks, e.g., AlexNet [19], VGG [54], ResNet [55]

TABLE 1
Recognition accuracies obtained by different methods on Volleyball Dataset.

Methods	Left pass	Right pass	Left set	Right set	Left spike	Right spike	Left win	Right win	Average
Ibrahim *et al.* [10]	77.9	81.4	84.5	68.8	89.4	85.6	88.2	87.4	82.9
Shu *et al.* [44]	-	-	-	-	-	-	-	-	83.6
Li *et al.* [58]	55.8	69.1	67.3	52.1	82.1	79.2	-	-	67.6
Biswas *et al.* [12]	-	-	-	-	-	-	-	-	83.0
Yan *et al.* [36]	85.8	88.1	**90.5**	**80.2**	92.2	87.9	89.2	90.8	88.1
B1 (frame-level CNN)	59.3	62.9	64.9	66.7	83.8	76.3	89.2	72.4	71.9
B2 (single-person CNN)	73.0	73.8	83.3	70.8	86.0	87.9	74.5	47.1	74.6
B3 (Graph LSTM)	**90.3**	84.8	87.5	79.2	91.1	88.4	86.3	85.1	86.6
B4 (with only TCC)	89.8	84.8	86.9	77.6	91.1	89.6	88.2	86.2	86.8
B5 (with only SCC)	89.8	83.8	87.5	78.7	91.6	89.0	87.3	87.4	86.9
B6 (with only STCC)	89.8	87.6	85.1	78.1	92.2	90.8	90.2	88.5	87.8
B7 (with only GCC)	88.5	87.1	85.7	77.1	88.8	**92.5**	**94.1**	86.2	87.5
CCG-LSTM	88.1	**90.0**	89.9	78.1	**93.9**	91.3	90.2	**93.1**	**89.3**

and GoogLeNet [56]. For fair comparison, we employ the pre-trained AlexNet model to extract the CNN feature of each person on the person's bounding box.

The CCG-LSTM is implemented with Pytorch toolbox on a NVIDIA Tesla K40 GPU. We use the Adam algorithm [57] with a learning rate of 0.0001 and a momentum of 0.99 for all networks to minimize the loss function, and the learning rate is decreased to $1/10$ of the original value after every five epochs.

4.3 Experiments on Volleyball Dataset

The Volleyball Dataset [10] consists of 55 videos with 4830 annotated frames, which are collected from YouTube. For every frame, the bounding box of each person is given, each person is labeled with one of the action classes (e.g. "Waiting", "Setting", "Digging", "Failing", "Spiking", "Blocking", "Jumping", "Moving" and "Standing"), and each video clip is labeled with one of the fine-grained activity classes (e.g. "Left pass", "Right pass", "Left set", "Right set", "Left spike", "Right spike", "Left win" and "Right win"). We split the training and testing sets following the same setting in [10], namely two-thirds of the annotated frames are used for training and the rest ones are used for testing. The dimensions of the input data of CCG-LSTM, the motion state of CCG-LSTM, and the hidden state of Aggregation LSTM are set to 4096, 3000 and 4096, respectively. In the Volleyball Dataset, there are two sub-groups corresponding to two teams of players. First, all the motion states of the individuals in each sub-group are aggregated into a representation via Aggregation LSTM. Second, following [10], we recognize the team activity based on the concatenation of the representations of two sub-group activities to avoid the confusion of "left" and "right".

Ablation studies. The recognition accuracy of the proposed CCG-LSTM compared with the baselines are shown in Table 1. The proposed CCG-LSTM achieves the best performance on average and some activity classes (e.g. "Right pass", "Left Spike", and "Right win"). Compared with B1 and B2, the variants (i.e., B3 – B7) of CCG-LSTM have significant improvements in recognition accuracy by jointly exploring the temporal motions of individuals and spatial interactions among persons. B4 and B5 are designed to illustrate the importance of TCC and SCC for capturing relevant motions in the temporal and spatial domains, respectively. In comparison to B3, the improvements obtained by B6 and B7 demonstrate that capturing the relevant motions by STCC and quantifying their different contributions by GCC are both effective for recognizing group activities. Since GCC measures the consistency between the individual action and the activity, it can better capture the difference between two same-type activities (such as "left win" vs. "right win") than STCC. Thus, B7 (with only GCC) gains better performance on some specific activities, such as "Left win". Either B6(STCC) or B7(with GCC) gains performance on some specific activities. However, compared with B6(STCC) and B7(with GCC), CCG-LSTM (with both STCC and GCC) obtains better and satisfactory performance on the activities of the right pass, left/right set, left spike and right win. Finally, CCG-LSTM gains the best average accuracy.

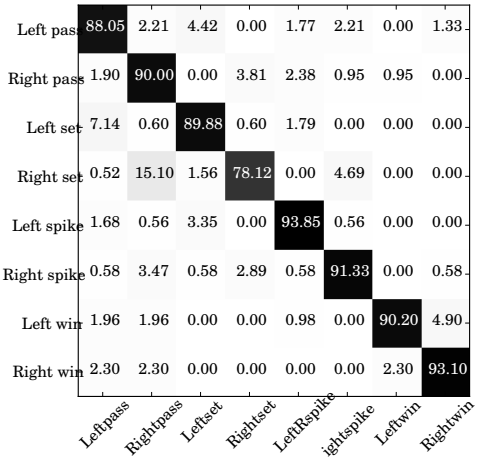


Fig. 6. Confusion matrix of CCG-LSTM on the Volleyball Dataset.

Comparison with the state-of-the-art methods. We compare the proposed CCG-LSTM with the state-of-the-art methods, including Ibrahim *et al.* [10], Shu *et al.* [44], Li *et al.* [58], Sovan *et al.* [12], and Yan *et al.* [36] on the Volleyball Dataset. Among these methods, Shu *et al.* [44] and Sovan *et al.* [12] do not provide the specific accuracy per class, and Li *et al.* [58] ignore the classes of "Left win" and "Right win". The recognition accuracies obtained by different methods are shown in Table 1. We can see that, the proposed CCG-LSTM achieves the best performance on average and in most of the classes. In particular, CCG-LSTM achieves approximately 6.4% improvement compared with the original work [10] that released this dataset. More im-

TABLE 2
Recognition accuracies obtained by different methods on CAD.

Methods	Moving	Waiting	Queuing	Talking	Average
Lan *et al.* [60]	92	69	76	99	84
Choi *et al.* [61]	90	**82.9**	95.4	94.9	90.8
Zhou *et al.* [62]	88.5	74.0	95.0	98.0	88.9
Ibrahim *et al.* [10]	95.9	66.4	96.8	**99.5**	89.7
Hajimirsadeghi *et al.* [59]	87	75	92	99	88.3
Wang *et al.* [42]	94.9	63.6	**100**	**99.5**	89.4
Li *et al.* [58]	90.8	81.4	99.2	84.6	89.0
Yan *et al.* [36]	92.8	76.6	**100**	**99.5**	92.2
B1 (frame-level CNN)	79.9	36.2	96.7	**99.5**	78.1
B2 (person-level CNN)	**97.6**	51.8	**100**	**99.5**	87.2
B3 (Graph LSTM)	96.2	38.3	**100**	**99.5**	83.5
B4 (with only TCC)	95.7	64.5	**100**	97.8	89.5
B5 (with only SCC)	95.2	55.3	**100**	**99.5**	87.5
B6 (with only STCC)	90.4	70.9	**100**	**99.5**	90.2
B7 (with only GCC)	94.3	63.1	**100**	**99.5**	89.2
CCG-LSTM	97.1	75.2	**100**	**99.5**	**93.0**

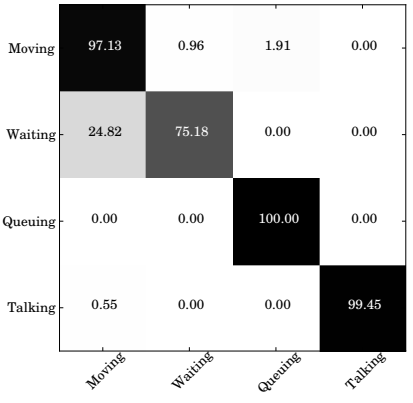


Fig. 7. Confusion matrix of the proposed CCG-LSTM on CAD.

portantly, the average performance of the proposed CCG-LSTM is better than Biswas *et al.* [12] that also models the individuals' motions and interactions jointly. These results demonstrate that the proposed CCG-LSTM is effective in modeling complex group activity among two sub-groups.

Confusion analysis. We analyze the confusion of the recognition result obtained by the proposed CCG-LSTM. The confusion matrix of the proposed CCG-LSTM on the Volleyball Dataset is shown in Figure 6. Since the actions of individuals are visually similar in most of the frames, group activity recognition in this dataset can be seen as a fine-grained recognition task. It is noted that the largest confusion happens between the "Right set" activity and "Right pass" activity, where there are many similar frames and fine-grained persons' interactions between these two types of activities. For example, the "Left pass" activity and the "Left set" activity are visually similar in most of the frames, and the difference of persons' interactions between the two types of activities is subtle in most of the frames.

4.4 Experiments on Collective Activity Dataset

The Collective Activity Dataset (CAD) [20] contains 44 video clips collected by a low-resolution hand-held camera. Each person is manually labeled with a fine-grained activity label, such as "Crossing", "Waiting", "Queuing", "Walking" or "Talking", and a pose label (not used in this work). Subsequently, each activity scene is assigned to a class label of group activity based on what the majority of people are doing in the scene. We follow the train/test split provided by [59], and use the person tracklets provided in [38]. Following the experimental settings in [42], we merge the classes "Walking" and "Crossing" into the class of "Moving" for fair comparison. The dimensions of the input of CCG-LSTM, the motion state of CCG-LSTM and the motion state of Aggregation LSTM are set to 4096, 2048 and 2048, respectively. Since the number of individuals varies from 1 to 12, we randomly select five effective persons for each frame and regard them as an entire group. If the number of persons in one frame is less than five, we take a full-zero matrix as the person's tracklets of a new person.

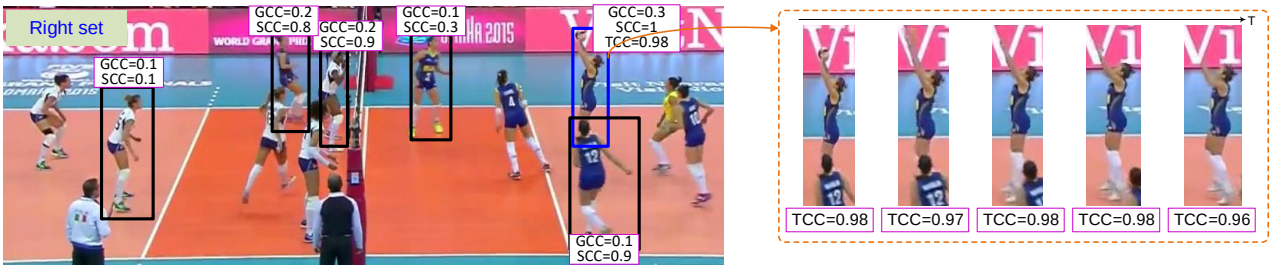
Ablation studies. The recognition accuracies obtained by the proposed CCG-LSTM and the baseline methods are illustrated in Table 2. The proposed CCG-LSTM obtains the best performance on average compared with all baselines. It is noted that B3, which considers the spatial interactions, has not improved the performance compared to B2, since there are few spatial interactions among persons in the "Waiting" activity. Likewise, B4 with TCC performs better than B5 with SCC, since the number of interacting persons is small. By considering STCC (consisting of TCC and SCC), the performance of B6 is improved compared with B1 – B5. If we jointly consider the STCC and GCC constraints, the average accuracy has been improved to 93.0%, corresponding to the best performance.

Comparison with the state-of-the-art methods. We compare the proposed CCG-LSTM with Lan *et al.* [60], Choi *et al.* [38], Zhou *et al.* [62], Ibrahim *et al.* [10], Hajimirsadeghi *et al.* [59], Wang *et al.* [42], Li *et al.* [58] and Yan *et al.* [36], and the recognition results obtained by different methods are shown in Table 2. The results of some methods [60], [38], [62], [10], [59], [42], [58], [36] are reported from the corresponding confusion matrices. As expected, the proposed CCG-LSTM obtains approximately 9% and 2% improvements compared with two previous non-deep learning methods (i.e., [60] and [61]), respectively. Not only that, CCG-LSTM is also better than the deep learning based methods, such as [62], [10], [59], [42], [58]. More importantly, the performance of the proposed CCG-LSTM is better than the state-of-the-art method [36], while it only performs worse on the "Waiting" activity due to the aforementioned issue.

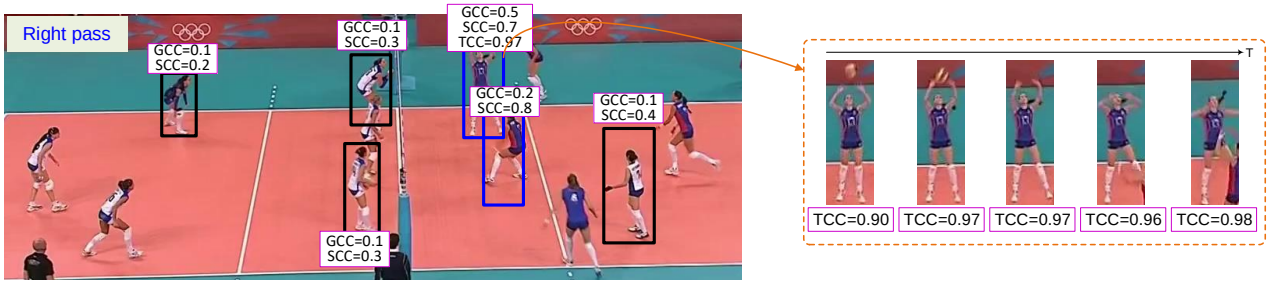
Confusion analysis. We show the confusion matrix of the proposed CCG-LSTM on Collective Activity Dataset in Figure 7. As mentioned before, the class of group activity is decided on what the majority of people are doing in this scene. For many activity classes, the interactions of individuals are visually similar in most of the frames. Thus, the group activity recognition on CAD is a fine-grained recognition task to some extent. It is noted that the "Waiting" activity is confused by the "Moving" activity seriously, since the difference of persons' interactions in these two types of activities are almost the same. In general, the proposed CCG-LSTM obtains satisfactory recognition accuracies for the fine-grained "Moving", "Queuing", and "Talking" activities.

4.5 Visualization Analysis

To illustrate the contributions of SCC, TCC, STCC, and GCC, we show some visualized examples by utilizing B4



(a) "Right set" activity in Volleyball Dataset.



(b) "Right pass" activity in Volleyball Dataset.

(c) "Moving" activity in CAD.

(d) "Talking" activity in CAD.

Fig. 8. Visualized examples of SCC, TCC, STCC, and GCC in modeling group activities. In each time step, one motion with the larger STCC Value (STCC_Value = SCC_Value × TCC_Value), as well as the larger GCC Value, it is more relevant to the activity.

(with TCC), B5 (with SCC), B6 (with STCC) and B7 (with GCC) in modeling group activities, as shown in Figure 8. However, visualizing the values of SCC, TCC, STCC, and GCC is not a trivial task in the modeling learning process. To this end, we adopt the following strategy to quantify them. First, we compute the TCC Value and SCC Value of the v-th person at time step t by TCC_Value $= mean(\tau_v^t)$, and SCC_Value $= mean(\boldsymbol{\varsigma}_v^t)$, where $mean(\cdot)$ denotes the average operation. Here, SCC_Value is further implemented by the Min-max Normalization. Based on the definition of STCC, the STCC Value of the v-th person at time step t is calculated by STCC_Value = SCC_Value × TCC_Value. For simplicity, the GCC Value of the v-th person at time step t is defined as GCC_Value $= \beta_v^t$, where all GCC Values in one frame are implemented by the 01 Normalization, respectively. Second, we mark the SCC Value, TCC Value, and GCC Value of each person at one time step. By this visualization method, one individual's motion with the larger STCC Value (STCC_Value = SCC_Value × TCC_Value), as well as the larger GCC Value in one time step, is more relevant to the group activity. For example, in Figure 8(a), one person in the blue bounding box is setting the volleyball, which is the most relevant to the "Right set" activity at this time step. We also find that this motion has the largest GCC Value (0.3), and the largest STCC Value ($1 \times 0.98 = 0.98$). In Figure 8(d), we can see that a person in the green bounding box does not participate in the " Talking" activity. We also find that this person has the smaller GCC Value (0.1), and

the smallest SCC Value (0.3). It can be concluded that STCC (a combination of SCC and TCC) and GCC can positively capture the relevant motions in the model learning process.

5 Conclusions

In this work, we explore the motion-level characteristics of group activity with several coherence constraints and propose a novel Coherence Constrained Graph LSTM (CCG-LSTM) for group activity recognition. Specifically, the extracted CNN features of persons are fed into the CCG-LSTM with STCC and GCC to capture the relevant motions, as well as quantify their different contributions. Subsequently, an Aggregation LSTM aggregates the motion states of all individuals into a hidden representation of the whole activity at a certain time step. Finally, all the hidden representations are input into the corresponding softmax classifiers, of which the outputs are combined to infer the class of the group activity. We conduct experiments on two widely-used datasets to demonstrate the effectiveness of the proposed method compared with the-state-of-art methods.

References

[1] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems (NIPS)*, 2014.

[2] X. Shu, J. Tang, G.-J. Qi, Y. Song, Z. Li, and L. Zhang, "Concurrence-aware long short-term sub-memories for person-person action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2017.

[3] H. Wang, D. Oneata, J. Verbeek, and C. Schmid, "A robust and efficient video representation for action recognition," *International Journal of Computer Vision*, vol. 119, no. 3, pp. 219–238, 2016.

[4] G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri, and B. P. Buckles, "Advances in human action recognition: A survey," *arXiv preprint arXiv:1501.05964*, 2015.

[5] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *IEEE International Conference on Computer Vision (ICCV)*, 2009.

[6] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[7] M. Ziaeefard and R. Bergevin, "Semantic human activity recognition: a literature review," *Pattern Recognition*, vol. 48, no. 8, pp. 2329–2345, 2015.

[8] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural Computation*, vol. 1, no. 2, pp. 270–280, 1989.

[9] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[10] M. Ibrahim, S. Muralidharan, Z. Deng, A. Vahdat, and G. Mori, "A hierarchical deep temporal model for group activity recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[11] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *Neural Information Processing Systems (NIPS)*, 2016.

[12] S. Biswas and J. Gall, "Structural recurrent neural network (srnn) for group activity analysis," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018.

[13] Z. Deng, A. Vahdat, H. Hu, and G. Mori, "Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] C. Zhang, X. Yang, J. Zhu, and W. Lin, "Parsing collective behaviors by hierarchical model with varying structure," in *ACM International Conference on Multimedia (ACM MM)*, 2012.

[15] Z. Deng, M. Zhai, L. Chen, Y. Liu, S. Muralidharan, M. J. Roshtkhari, and G. Mori, "Deep structured models for group activity recognition," in *British Machine Vision Conference (BMVC)*, 2015.

[16] Y. Li, D. Tarlow, M. Brockschmidt, and R. Zemel, "Gated graph sequence neural networks," *arXiv preprint: 1511.05493*, 2015.

[17] V. Ramanathan, J. Huang, S. Abu-El-Haija, A. Gorban, K. Murphy, and L. Fei-Fei, "Detecting events and key actors in multi-person videos," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[18] Y. Tang, L. Ma, W. Liu, and W. Zheng, "Long-term human motion prediction by modeling motion context and enhancing motion dynamic," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Neural Information Processing Systems (NIPS)*, 2012.

[20] W. Choi, K. Shahid, and S. Savarese, "What are they doing?: Collective activity classification using spatio-temporal relationship among people," in *IEEE International Conference on Computer Vision (ICCV) Workshops*, 2009.

[21] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[22] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[23] X. Chang, W.-S. Zheng, and J. Zhang, "Learning person–person interaction in collective activity recognition," *IEEE Transactions on Image Processing*, vol. 24, no. 6, pp. 1905–1918, 2015.

[24] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 886–893.

[25] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *ACM International Conference on Multimedia (ACM MM)*, 2007.

[26] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[27] I. Sutskever, J. Martens, and G. E. Hinton, "Generating text with recurrent neural networks," in *International Conference on Machine Learning (ICML)*, 2011.

[28] M. Freitag, S. Amiriparian, S. Pugachevskiy, N. Cummins, and B. Schuller, "audeep: Unsupervised learning of representations from audio with deep recurrent neural networks," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 6340–6344, 2017.

[29] 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." 2016.

[30] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[31] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *IEEE International Conference on Computer Vision (ICCV)*, 2015.

[32] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[33] Z. Wu, Y.-G. Jiang, X. Wang, H. Ye, and X. Xue, "Multi-stream multi-class fusion of deep networks for video classification," in *ACM International Conference on Multimedia (ACM MM)*, 2016.

[34] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+d: A large scale dataset for 3d human activity analysis," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[35] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

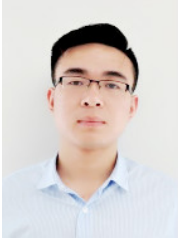
[36] R. Yan, J. Tang, X. Shu, Z. Li, and Q. Tian, "Participation-contributed temporal dynamic model for group activity recognition," in *ACM Conference on Multimedia (MM)*, 2018.

[37] W. Lin, Y. Chen, J. Wu, H. Wang, B. Sheng, and H. Li, "A new network-based algorithm for human activity recognition in

videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 5, pp. 826–841, 2014.
[38] W. Choi and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," in *European Conference on Computer Vision (ECCV)*, 2012.
[39] T. M. 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
[40] J. Yang, M. N. Nguyen, P. P. San, X. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.
[41] W. Lin, H. Chu, J. Wu, B. Sheng, and Z. Chen, "A heat-map-based algorithm for recognizing group activities in videos," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 11, pp. 1980–1992, 2013.
[42] M. Wang, B. Ni, and X. Yang, "Recurrent modeling of interaction context for collective activity recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
[43] Y. Zhang, X. Liu, M. Chang, W. Ge, and T. Chen, "Spatio-temporal phrases for activity recognition," in *European Conference on Computer Vision (ECCV)*, 2012.
[44] T. Shu, S. Todorovic, and S.-C. Zhu, "Cern: confidence-energy recurrent network for group activity recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
[45] W. Lin, M.-T. Sun, R. Poovendran, and Z. Zhang, "Group event detection with a varying number of group members for video surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 20, no. 8, pp. 1057–1067, 2010.
[46] B. Jiang, H. Chen, B. Yuan, and X. Yao, "Scalable graph-based semi-supervised learning through sparse bayesian model," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 12, pp. 2758–2771, 2017.
[47] J. Tang, X. Shu, Z. Li, Y.-G. Jiang, and Q. Tian, "Social anchor-unit graph regularized tensor completion for large-scale image retagging," *arXiv preprint: 1804.04397*, 2018.
[48] Y. Yuan, X. Liang, X. Wang, D.-Y. Yeung, and A. Gupta, "Temporal dynamic graph lstm for action-driven video object detection." in *IEEE International Conference on Computer Vision (ICCV)*, 2017.
[49] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan, "Semantic object parsing with graph lstm," in *European Conference on Computer Vision (ECCV)*, 2016, pp. 125–143.
[50] Y. Li, R. Zemel, M. Brockschmidt, and D. Tarlow, "Gated graph sequence neural networks," in *International Conference on Learning Representations (ICLR)*, 2016.
[51] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint: 1409.0473*, 2014.
[52] M. Danelljan, G. Häger, F. Khan, and M. Felsberg, "Accurate scale estimation for robust visual tracking," in *British Machine Vision Conference (BMVC)*, 2014.
[53] D. E. King, *Dlib-ml: A machine learning toolkit*, 2009, vol. 10, no. 7.
[54] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, 2015.
[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
[56] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.
[57] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.
[58] X. Li and M. C. Chuah, "SBGAR: semantics based group activity recognition," in *IEEE International Conference on Computer Vision (ICCV)*, 2017.
[59] H. Hajimirsadeghi, W. Yan, A. Vahdat, and G. Mori, "Visual recognition by counting instances: A multi-instance cardinality potential kernel," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2015.
[60] 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*, vol. 34, no. 8, pp. 1549–1562, 2012.
[61] W. Choi and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," in *European Conference on Computer Vision (ECCV)*, 2012.
[62] Z. Zhou, K. Li, X. He, and M. Li, "A generative model for recognizing mixed group activities in still images," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.

Jinhui Tang (M'08-SM'14) received the BEng and PhD degrees from the University of Science and Technology of China, in 2003 and 2008, respectively. He is currently a professor with the Nanjing University of Science and Technology. He has authored more than 150 papers in top-tier journals and conferences. His research interests include multimedia analysis and computer vision. He was a recipient of the best paper awards in ACM MM 2007, PCM 2011 and ICIMCS 2011, the Best Paper Runner-up in ACM MM 2015, and the best student paper awards in MMM 2016 and ICIMCS 2017. He has served as an associate editor of the IEEE TNNLS, the IEEE TKDE, and the IEEE TCSVT. He is a senior member of the IEEE.

Xiangbo Shu is an Associate Professor in School of Computer Science and Engineering, Nanjing University of Science and Technology, China. He received his Ph.D. degree in July 2016 from Nanjing University of Science and Technology. From 2014 to 2015, he worked as a visiting scholar in the Department of Electrical and Computer Engineering at National University of Singapore. His research interests include computer vision, multimedia computing and deep learning. He has received the Excellent Doctoral Dissertation of CAAI, the Excellent Doctoral Dissertation of Jiangsu Province, the Best Student Paper Award in MMM 2016 and the Best Paper Runner-up in ACM MM 2015. He is a member of the IEEE, ACM, and CCF.



Rui Yan is currently a PhD candidate of School of Computer Science and Engineering, Nanjing University of Science and Technology, China. Before that, he received the BEng degree from the Nanjing Forestry University, China, in 2017. His research interests include computer vision, and deep learning. He received the Excellent Undergraduate Thesis of Jiangsu Province.

Liyan Zhang received the Ph.D. degree in computer science from the University of California, Irvine, in 2014. She is currently a Professor with the School of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, China. Her research interests include multimedia analysis, and computer vision. She has received the Best Paper Award in ICMR 2013 and the Best Student Paper Award in MMM 2016.