Hierarchical Long Short-Term Concurrent Memory for Human Interaction Recognition

Xiangbo Shu, Jinhui Tang, *Senior Member, IEEE,* Guo-Jun Qi, Wei Liu and Jian Yang

***Abstract*—In this work, we aim to address the problem of human interaction recognition in videos by exploring the long-term inter-related dynamics among multiple persons. Recently, Long Short-Term Memory (LSTM) has become a popular choice to model individual dynamic for single-person action recognition due to its ability to capture the temporal motion information in a range. However, most existing LSTM-based methods focus only on capturing the dynamics of human interaction by simply combining all dynamics of individuals or modeling them as a whole. Such methods neglect the inter-related dynamics of how human interactions change over time. To this end, we propose a novel Hierarchical Long Short-Term Concurrent Memory (H-LSTCM) to model the long-term inter-related dynamics among a group of persons for recognizing human interactions. Specifically, we first feed each person's static features into a Single-Person LSTM to model the single-person dynamic. Subsequently, at one time step, the outputs of all Single-Person LSTM units are fed into a novel Concurrent LSTM (Co-LSTM) unit, which mainly consists of multiple sub-memory units, a new cell gate, and a new co-memory cell. In the Co-LSTM unit, each sub-memory unit stores individual motion information, while this Co-LSTM unit selectively integrates and stores inter-related motion information between multiple interacting persons from multiple sub-memory units via the cell gate and co-memory cell, respectively. Extensive experiments on several public datasets validate the effectiveness of the proposed H-LSTCM by comparing against baseline and state-of-the-art methods.**

***Index Terms*—Human interaction recognition, long short-term memory, activity recognition, deep learning.**

I. Introduction

HUMAN interactions (e.g., handshaking, and talking) are typical human activities that occur in public places and are attracting substantial attention from researchers [1]–[4]. A human interaction usually involves at least two individual dynamics from multiple persons, who are concurrently inter-related with each other (e.g., some persons are talking together, some persons are handshaking with each other). In most cases of human interaction, the concurrent inter-related dynamics between multiple persons are strongly interacting (e.g., person A kicks person B, while person B retreats back). It has been shown that the concurrent inter-related dynamics among multiple persons, rather than single-person dynamics, can contribute discriminative information for recognizing human interactions [5].

Two main types of solutions exist for the problem of human interaction recognition. One solution (e.g., [1], [2], [6], [7]) is to extract the dynamic descriptors from each interacting person, and then predict the class of human interaction by inferring the coherence between two individual dynamics. However, this solution, i.e., regarding human interactions as multiple single-person actions, ignores some inter-related motion information and brings in some irrelevant individual motion information. The other solution is to extract motion descriptors of interacting regions, and then to train an interaction recognition model [5]. However, interacting regions are difficult to locate before the close interaction occurs.

X. Shu, J. Tang and J. Yang are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China. E-mail: {shuxb, jinhuitang, csjyang}@njust.edu.cn. (Corresponding author: Jinhui Tang)

G.-J. Qi is with the Department of Electrical Engineering and Computer Science, University of Central Florida, Orlando, Florida 32816, USA. Email: guojun.qi@ucf.edu

W. Liu is with the Computer Vision Group, Tencent AI Lab, Shenzhen 518000, China. E-mail: wliu@ee.columbia.edu

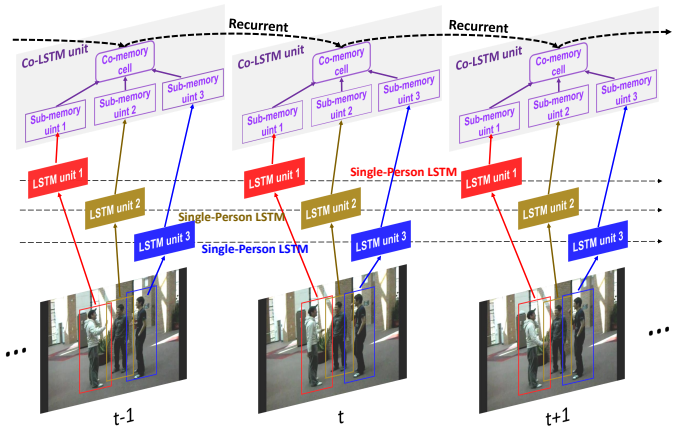


Fig. 1. The framework of the proposed Hierarchical Long Short-Term Concurrent Memory (H-LSTCM) for modeling human interactions in a human interaction scene. The details of Co-LSTM unit is displayed in Figure 2.

Recently, due to the powerful ability to capture sequential motion information, Long Short-Term Memory (LSTM) [8] has proven to be successful for various human action recognition tasks [9]–[13]. Therefore, we consider exploring the long-term inter-related dynamics among a group of interacting persons by LSTM. However, conventional LSTM models human dynamics independently, and does not consider the concurrent inter-relation of dynamics among multiple persons. A straightforward way to overcome this limitation is to either 1) merge individual actions at the preprocessing stage [14] (e.g., consider interacting persons as a whole); or 2) utilize several LSTMs to model the single-person dynamics of individuals respectively, and then fuse the output sequences of these LSTMs [13]. However, both methods neglect the inter-related dynamics of how interactions among persons change over time.

To this end, we propose a novel Hierarchical Long Short-Term Concurrent Memory (H-LSTCM) for human interaction (or human activity) recognition by modeling the long-term inter-related dynamics among a group of interacting persons, as shown in Figure 1. Specifically, for each person, we first feed her/his static features (e.g., CNN features) into a Single-Person LSTM to model the single-person dynamic, which describes a person's long-term motion information in a whole video clip. Then, all outputs of Single-Person LSTM units are fed into a novel Concurrent LSTM (Co-LSTM) unit, which mainly consists of multiple sub-memory units, multiple new cell gates, and a new co-memory cell. In a Co-LSTM unit, multiple sub-memory units store the single-person motion information from the Single-Person LSTM units. Following these sub-memory units, the cell gates allow the inter-related motion memory in sub-memory units to enter a new co-memory cell, which selectively integrates and stores the inter-related memory to reveal the concurrent inter-related motion information among persons. Overall, all interacting persons at each time step are jointly modeled by a Co-LSTM unit on the person's bounding boxes. At the last time step, the output

of Co-LSTM is a dynamic inter-related representation of the human activity. Extensive experiments on various datasets are conducted to evaluate the performance of H-LSTCM compared with the state-of-the-art methods.

The main contributions of this work are summarized as follows:

- We propose a novel Hierarchical Concurrent Long Short-Term Concurrent Memory (H-LSTCM) to effectively address the problem of human interaction recognition with multiple persons, by learning the dynamic inter-related representations among all persons in the human interaction scenes.
- We design a novel Concurrent LSTM (Co-LSTM) to aggregate the inter-related memory from individuals in human activity scenes over time, by capturing the concurrently long-term inter-related dynamics among multiple persons rather than the dynamics of individuals.

Our preliminary Co-LSTSM in [15] consists of two sub-memory units at each time step. In the two persons' interaction scene, there is only one person-person interaction between these two persons. Thus, Co-LSTSM can directly learn the inter-related representation of the two persons' interaction. However, it cannot handle the multiple ($\geq$ 3) persons' interactions. In this work, the proposed H-LSTCM (by stacking Single-Person LSTM and Co-LSTM in a hierarchical way) can recognize different kinds of interactions among persons, two persons' interaction, multiple ($\geq$ 3) persons interaction, collective activity with multiple ($\geq$ 3) persons, and group activity with multiple sub-group activities. Moreover, Co-LSTSM learns the dynamic inter-related representation between two persons simply from the static single-person features. Actually, there is a large gap between the static single-person features and the dynamic inter-related representation, which limits the performance of the Co-LSTSM. Thus, in H-LSTCM, we bring in the single-person dynamic, which is a basic element in the multiple persons' interaction or group activity to describe a person's long-term motion information in a whole video clip, and reflects motion patterns caused by interactions with other persons. H-LSTCM learns the dynamic inter-related representation among multiple persons in a hierarchical way, from the static to dynamic features at the single-person level first, and further to an inter-related level of group activities. Specifically, the single-person LSTMs in H-LSTCM first learn dynamic single-person features of persons from the static single-person features. And then, an extended Co-LSTM with multiple sub-memory units in H-LSTCM learns concurrently inter-related representation among all persons based on the single-person dynamics. Such a hierarchical strategy ensures that H-LSTCM learns more discriminative representation than Co-LSTSM for multiple persons' interactions.

II. Related Work

Human interaction/activity recognition aims to automatically understand the interaction performed by at least two persons in a scene [2]. In the task of two persons' interaction (or person-person interaction) recognition, earlier researchers have noted that the interactive attributes of persons provide discriminative information to represent person-person interactions. For example, Kong *et al.* [1], [6] regarded multiple interactive phrases as the latent mid-level feature to recognize person-person interactions from human individual actions. Consider the temporal context information in videos, Zhang *et al.* [7] and Liu *et al.* [16] used a new set of spatio-temporal action attribute phrases to describe the two persons' interactions. However, the difference in some person-person interactions (e.g., boxing and patting) is too small to be identified via only interactive phrases. Moreover, some complex person-person interactions cannot be described well by a specified number of interactive phrases.

Benefiting from the success of deep learning, some deep learning-based methods have been proposed to understand person-person interactions for the past few years [14], [17]. For example, Wang *et al.* [17] adopted the deep context features instead of the traditional context features (e.g., [18]) to recognize person-person interactions. One limitation of the above methods is that locating the interactive region is a challenging task before the close interaction occurs. Therefore, this work aims to recognize human interactions without locating the interactive regions accurately.

In a scene of multiple persons' interaction (i.e., human activity), several persons interact with each other, which makes activity recognition complex. Two solutions are commonly used to address the problem of human activity recognition. One solution is to exploit the spatio-temporal descriptors to represent the spatio-temporal distribution of persons in the human activity [19]–[21]. The other solution is to track all the body parts in the video, and then learn holistic representations to estimate the class of the human activity [22], [23]. However, the former solution requires inference of the complex spatio-relation between persons, and the latter one brings in some individual motion information of outlier persons.

Recently, Long Short Term Memory (LSTM) has been proposed to address the problem of human interaction recognition by learning high-level dynamic representations of persons [13], [14]. This insight motivates us to employ LSTM to learn high-level dynamic representations of human activity. However, traditional LSTM targeting single-person actions cannot handle multiple-person interactions well. As mentioned previously, we can roughly treat all the interacting persons as a whole for training LSTM. However, this solution results in some individual-specific motion information. Additionally, we can model the single-person dynamics of individuals by multiple LSTMs, respectively, and then integrate (e.g., concatenate or pool) the single-person dynamics obtained by all these LSTMs into the final representation. Since this strategy assumes that all persons in a human activity are independent of each other, the crucial inter-related motion information among persons is lost. Therefore, we propose a new Hierarchical Long Short-Term Concurrent Memory (H-LSTCM) that adopts a hierarchical way to first model the single-person dynamics of individuals by several Single-Person LSTMs, and then model the inter-related dynamics among persons by a new Co-LSTM.

Closely related work includes Hierarchical Deep Temporal Model (HDTM) [13], Deep Structured Model (DSM) [24], and Structure Inference Machines (SIM) [25]. Specifically, HDTM [13] first models the individual dynamics by several LSTMs. And then, the outputs of these LSTMs are pooled into a single vector, which is the input of the following LSTM. HDTM pools single-person dynamics into an overall dynamic representation, which ignores the inter-relations among persons in the human activity. DSM [24] and SIM [25] utilize CNN to obtain the initialized class labels of single-person actions and group-level activity, and refine the group activity class label by exploring the inter-relations among the actions of all individuals in an iterative manner. If one person's action is closely related to the group activity and the other persons' actions, this person intensively participates in the group activity; otherwise, this person is an outlier. Since DSM and SIM target "key" persons who play crucial roles in the group activity rather than all persons, some sudden motion information of "outlier" persons may be lost. Actually, we observe that outlier persons are not irrelevant to the group activity at any time step. For example, a person may suddenly spike the ball at one moment in a volleyball game. Overall, compared to HDTM [13], the proposed H-LSTCM considers the inter-relations among persons via the cell gates and co-memory cell. And compared to DSM [24] and SIM [25], H-LSTCM models the concurrently inter-related dynamics among all persons, which include the outlier yet useful persons.

It is noted that some works [26], [27] proposed to learn the location-related representation among multiple persons for multi-target tracking. They assume that two persons who have the close position are inter-related with each other. By contrast, the proposed H-LSTCM learns semantic-related representation among multiple persons by leveraging the inter-relation between the single-person dynamic at the current time step and the dynamic of the whole activity at the previous time step. Here, it is assumed that one person, whose current representation is closely related to the hidden representation of the whole activity, is likely to be more involved in this activity.

III. Preliminary

Given a video clip $\{\mathbf{x}_t \in \mathbb{R}^n | t = 1, \cdots, T\}$ with length T, where $\mathbf{x}_t$ is the static feature at time step t, a traditional Recurrent Neural Network (RNN) [28] models the dynamics of this video clip through a sequence of hidden states. Due to the exponential decay in retaining the context information of video frames [8], RNN does not model the long-term dynamics of video sequences well. To this end, Long Short-Term Memory (LSTM) [8], a variant of the RNN, provides a solution by incorporating memory units that enable the network to learn when to forget previous hidden states and when to update hidden states given new information [9].

A traditional LSTM unit [8] at time step t contains an input gate $\mathbf{i}_t$, a forget gate $\mathbf{f}_t$, an output gate $\mathbf{o}_t$, and a memory cell $\mathbf{c}_t$, which are expressed as follows,

$$\mathbf{i}_t = \sigma(\mathbf{W}_{ix} \cdot \mathbf{x}_t + \mathbf{W}_{ih} \cdot \mathbf{h}_{t-1} + \mathbf{b}_i); \quad (1)$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{fx} \cdot \mathbf{x}_t + \mathbf{W}_{fh} \cdot \mathbf{h}_{t-1} + \mathbf{b}_f); \quad (2)$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_{ox} \cdot \mathbf{x}_t + \mathbf{W}_{oh} \cdot \mathbf{h}_{t-1} + \mathbf{b}_o); \quad (3)$$

$$\mathbf{g}_t = \varphi(\mathbf{W}_{gx} \cdot \mathbf{x}_t + \mathbf{W}_{gh} \cdot \mathbf{h}_{t-1} + \mathbf{b}_g); \quad (4)$$

$$\mathbf{c}_t = \mathbf{f}_t^s \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t, \quad (5)$$

where $\sigma(\cdot)$ is a sigmoid function, $\odot$ denotes an element-wise product, $\varphi(\cdot)$ is a hyperbolic tangent $tanh(\cdot)$, $\mathbf{W}_{*x}$ and $\mathbf{W}_{*h}$ are weight matrices, and $\mathbf{b}_*$ is a bias vector. Subsequently, a hidden state $\mathbf{h}_t$ at time step t can be expressed as

$$\mathbf{h}_t = \mathbf{o}_t \odot \varphi(\mathbf{c}_t), \quad (6)$$

which denotes the dynamic representation of the t-th frame. All hidden states $\{\mathbf{h}_t | t = 1, 2, \cdots, T\}$ describe the dynamic of the video clip. Finally, the output $\mathbf{z}_t \in \mathbb{R}^k$ at time step t is computed as

$$\mathbf{z}_t = \varphi(\mathbf{W}_{zh} \cdot \mathbf{h}_t + \mathbf{b}_z), \quad (7)$$

which can be transformed to a probability $y_{t,l}$ ($l = 1, \cdots, k$) corresponding to the l-th class of the human activity by a softmax function

$$y_{t,l} = \frac{\exp(z_{t,l})}{\sum_{j=1}^{k} \exp(z_{t,j})}, \quad (8)$$

where $z_{t,j}$ in $\mathbf{z}_t$ denotes the encoding of the confidence score on the j-th activity class. Generally, we set $\mathbf{y}_t = [y_{t,1}, y_{t,2}, \cdots, y_{t,k}]^T$ as the predicted class label vector.

IV. Hierarchical Long Short-Term Concurrent Memory

A. The Architecture

For a video describing a human interaction (or human activity), each frame contains at least two concurrent singe-person actions of individuals, which are inter-related in this human interaction. In this work, we propose a Hierarchical Long Short-Term Concurrent Memory (H-LSTCM) to capture the concurrently inter-related dynamics among all the persons. Specifically, the proposed H-LSTCM first models the single-person dynamics of persons by multiple Single-Person LSTMs corresponding to these persons, and then captures the inter-related dynamics among all the persons by a new Concurrent LSTM (Co-LSTM). Figure 1 shows the whole framework of the proposed H-LSTCM. The key point of H-LSTCM is to utilize multiple sub-memory units in a Concurrent LSTM (Co-LSTM) unit to selectively integrate and store the concurrently inter-related motion information among multiple persons from the individual motions information.

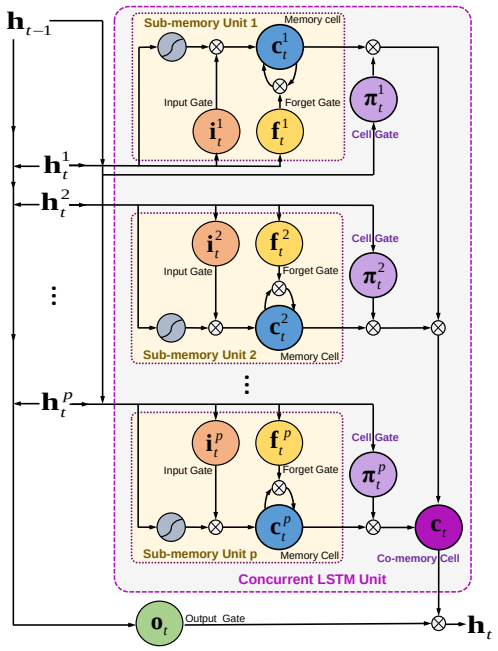


Fig. 2. Illustration of a Concurrent LSTM (Co-LSTM) unit in H-LSTCM.

Figure 2 illustrates the architecture of a Co-LSTM unit of the proposed H-LSTCM at a time step. The Co-LSTM unit mainly consists of multiple specific sub-memory units (the number of units corresponds to the number of interacting persons), multiple cell gates, a common output gate, and a new co-memory cell. Specifically, all sub-memory units include their respective input gates, forget gates, and memory cells. Following these sub-memory units, the cell gates allow the inter-related motion memory in the sub-memory units to enter a new co-memory cell, and the co-memory cell selectively integrates and memorizes the inter-related motion information among persons. Overall, the stacked Co-LSTM units are recurrent in a time sequence to capture the concurrently inter-related dynamics among all interacting persons over time.

Formally, let $\{\mathbf{x}_t^1 \in \mathbb{R}^n | t = 1, \cdots, T\}$, $\{\mathbf{x}_t^2 \in \mathbb{R}^n | t = 1, \cdots, T\}$, $\cdots$, and $\{\mathbf{x}_t^p \in \mathbb{R}^n | t = 1, \cdots, T\}$ denote the sets of static features (e.g., CNN features) on the person's bounding box (obtained by object detector and object tracker) of all p interacting persons in a video clip v_n (wherein $n = 1, 2, \cdots, N$). For a feature set of $\{\mathbf{x}_t^s | t = 1, \cdots, T\}$ of the s-th person, we can obtain her/his single-person hidden state (i.e., single-person dynamic) $\{\mathbf{h}_t^s | t = 1, \cdots, T\}$ at each time step via a Single-Person LSTM. In the s-th sub-memory unit of Co-LSTM on the top of the Single-Person LSTMs, $\mathbf{i}_t^s$, $\mathbf{f}_t^s$, $\mathbf{g}_t^s$, and $\mathbf{c}_t^s$ ($s = 1, 2, \cdots, p$) denote the input gate, forget gate, input modulation gate and sub-memory cell at time step t, respectively. These components can be expressed by the following equations

$$\mathbf{i}_t^s = \sigma(\mathbf{W}_{ix}^s \cdot \mathbf{h}_t^s + \mathbf{W}_{ih}^s \cdot \mathbf{h}_{t-1} + \mathbf{b}_i^s); \quad (9)$$

$$\mathbf{f}_t^s = \sigma(\mathbf{W}_{fx}^s \cdot \mathbf{h}_t^s + \mathbf{W}_{fh}^s \cdot \mathbf{h}_{t-1} + \mathbf{b}_f^s); \quad (10)$$

$$\mathbf{g}_t^s = \varphi(\mathbf{W}_{gx}^s \cdot \mathbf{h}_t^s + \mathbf{W}_{gh}^s \cdot \mathbf{h}_{t-1} + \mathbf{b}_g^s); \quad (11)$$

$$\mathbf{c}_t^s = \mathbf{f}_t^s \odot \mathbf{c}_{t-1}^s + \mathbf{i}_t^s \odot \mathbf{g}_t^s, s=1,2,\cdots,p, \quad (12)$$

where W_{*x}^s and W_{*h}^s are weight matrices, b_* is a bias vector, and the common hidden state $\mathbf{h}_{t-1}$ (defined in Eq. (16)) denotes the dynamic inter-related representation of the whole activity at time step $(t-1)$. All hidden states $\{\mathbf{h}_t | t = 1, 2, \cdots, T\}$ describe the inter-related dynamic of the activity scene in the video clip.

Following the s-th sub-memory unit, a new cell gate π_t^s aims to allow the inter-related memory in the s-th memory cell to enter the co-memory cell at time step t. Similar to traditional gates, the cell gate π_t^s is activated by a nonlinear function of the input $\mathbf{h}_t^s$ and the past hidden state $\mathbf{h}_{t-1}$ of the whole activity,

$$\pi_t^s = \sigma(\mathbf{W}_{\pi \mathrm{h}}^s \cdot \mathbf{h}_t^s + \mathbf{W}_{\pi h} \cdot \mathbf{h}_{t-1} + \mathbf{b}_\pi), s \in \{1, 2, \cdots, p\}, \quad (13)$$

where $\mathbf{W}_{\pi h}^s$ and $\mathbf{W}_{\pi h}$ are the weight matrices, and $\mathbf{b}_\pi$ is the bias vector. Based on the consistent interactions among multiple interacting persons, all cell gates π_t^s $(s = 1, 2, \cdots, p)$ allow more concurrently inter-related motion information among persons to enter a new co-memory cell $\mathbf{c}_t$, which contributes to a common hidden state $\mathbf{h}_t$ at time step t. In this work, the co-memory cell $\mathbf{c}_t$ can be expressed as

$$\mathbf{c}_t = \sum_{s=1}^{p} \pi_t^s \odot \mathbf{c}_t^s. \quad (14)$$

This co-memory cell $\mathbf{c}_t$ corresponds to an output gate $\mathbf{o}_t$ that is related to all the inputs and the common hidden state at the previous time step, i.e.,

$$\mathbf{o}_t = \sigma(\sum_{s=1}^{p} \mathbf{W}_{\mathrm{ox}}^s \mathbf{h}_t^s + \mathbf{W}_{oh} \cdot \mathbf{h}_{t-1} + \mathbf{b}_o). \quad (15)$$

Finally, the hidden state $\mathbf{h}_t$ at time step t can be expressed as

$$\mathbf{h}_t = \mathbf{o}_t \odot \varphi(\mathbf{c}_t). \quad (16)$$

If we obtain $\mathbf{h}_t$, we can compute the probability vector $\mathbf{y}_t$ of one human interaction by Eq (7) and Eq (8).

B. Learning Algorithm

We employ a loss function to learn the model parameters of H-LSTCM by measuring the deviation between the ground-truth class label vector $\hat{\mathbf{y}}_t = [\hat{y}_{t,1}, \hat{y}_{t,2}, \cdots, \hat{y}_{t,k}]^T$ and the predicted probability vector $\mathbf{y}_t = [y_{t,l}, y_{t,2}, \cdots, y_{t,k}]^T$ corresponding to $\mathbf{h}_t$ at time step t,

$$\ell(\mathbf{y}_t, l) = -\sum_{l=1}^{k} \hat{y}_{t,l} \log y_{t,l}. \quad (17)$$

When the training label of the activity frame at time step t corresponds to the target class l_t ($l_t \in \{1, 2, \cdots, k\}$), one element $\hat{y}_{t,l_t}$ in $\hat{\mathbf{y}}_t$ is set $\hat{y}_{t,l_t} = 1$, and the other elements in $\hat{\mathbf{y}}_t$ are zero. Then, Eq. (17) can be simplified as

$$\ell(\mathbf{y}_t, l_t) = -\log y_{t,l_t}, \quad (18)$$

where y_{t,l_t} is defined in Eq. (8). Some researchers [8], [11] have indicated that the memory cell of LSTM at the last time step can store useful sequence information of the whole data sequence (e.g., a video clip). That is, for a video clip of length T, if its class label l is annotated at the video level, the H-LSTCM model can be trained by minimizing the loss at time step T, i.e., $\ell(\mathbf{y}_T, l) = -\log y_{T,l}$. Otherwise, if the class label l is annotated on each frame t, we can minimize the cumulative loss over the sequence, i.e., $\sum_{t=1}^{T} \ell(\mathbf{y}_t, l)$.

In this work, given a training video clip with label l ($l \in \{1, 2, \cdots, k\}$) at the video level, we choose the loss

$$\mathcal{J}(\Theta) = \ell(\mathbf{y}_T, l), \quad (19)$$

Algorithm 1 Training for H-LSTCM

Input: N video clips, $Epoch$, Configuration set of H-LSTCM.
Initialization: Parameter set Θ, $epoch \leftarrow 1$.

1: Extract fc6 features of each person on the detected bounding box in each frame of each video.
// Forward propagation
2: Forward propagation of Single-person LSTMs;
3: Forward propagation of Co-LSTM.
// Back propagation
4: **for** $epoch = 1, 2, \cdots, Epoch$ **do**
5: **for** $n = 1, 2, \cdots, N$ **do**
6: Update parameters in Single-Person LSTMs via BPTT;
7: $\mathbf{b}_z \leftarrow \left(m \cdot \mathbf{b}_o^s - \eta \cdot \frac{\partial \mathcal{J}(\Theta)}{\partial \mathbf{b}_z}\right)$ [1];
8: $\mathbf{W}_{zh} \leftarrow \left(m \cdot \mathbf{b}_o^s - \eta \cdot \frac{\partial \mathcal{J}(\Theta)}{\partial \mathbf{W}_{zh}}\right)$;
9: $\mathbf{b}_o \leftarrow \left(m \cdot \mathbf{b}_o^s - \eta \cdot \frac{\partial \mathcal{J}(\Theta)}{\partial \mathbf{b}_o}\right)$;
10: $\mathbf{W}_{oh} \leftarrow \left(m \cdot \mathbf{W}_{oh} - \eta \cdot \frac{\partial \mathcal{J}(\Theta)}{\partial \mathbf{W}_{oh}}\right)$;
11: $\mathbf{W}_{ox}^s \leftarrow \left(m \cdot \mathbf{W}_{ox}^s - \eta \cdot \frac{\partial \mathcal{J}(\Theta)}{\partial \mathbf{W}_{ox}^s}\right)$, $s = 1, \cdots, p$;
12: $\mathbf{b}_\pi \leftarrow \left(m \cdot \mathbf{b}_\pi - \eta \cdot \frac{\partial \mathcal{J}(\Theta)}{\partial \mathbf{b}_\pi}\right)$;
13: $\mathbf{W}_{\pi h}^s \leftarrow \left(m \cdot \mathbf{W}_{\pi h}^s - \eta \cdot \frac{\partial \mathcal{J}(\Theta)}{\partial \mathbf{W}_{\pi h}^s}\right)$;
14: $\mathbf{W}_{\pi h} \leftarrow \left(m \cdot \mathbf{W}_{\pi h} - \eta \cdot \frac{\partial \mathcal{J}(\Theta)}{\partial \mathbf{W}_{\pi h}}\right)$;
15: $\mathbf{b}_*^s \leftarrow \left(m \cdot \mathbf{b}_*^s - \eta \cdot \frac{\partial \mathcal{J}(\Theta)}{\partial \mathbf{b}_*^s}\right)$, and $* \in \{i, f, g\}$;
16: $\mathbf{W}_{*x}^s \leftarrow \left(m \cdot \mathbf{W}_{*x}^s - \eta \cdot \frac{\partial \mathcal{J}(\Theta)}{\partial \mathbf{W}_{*x}^s}\right)$;
17: $\mathbf{W}_{*h}^s \leftarrow \left(m \cdot \mathbf{W}_{*h}^s - \eta \cdot \frac{\partial \mathcal{J}(\Theta)}{\partial \mathbf{W}_{*h}^s}\right)$.
18: **end for**
19: **end for**

Output: Parameter set Θ.

[1] Here, m and θ are the momentum parameter and learning rate, respectively. The detailed deductions of the derivative of all the parameters can be found in Appendix A.

where Θ denotes a parameter set including all the parameters of the H-LSTCM model. The loss function of H-LSTCM can be minimized by Backpropagation Through Time (BPTT). The detailed deductions of the derivatives of all the parameters in the H-LSTCM model can be found in Appendix A of the supplemental material. The detailed training procedure of H-LSTCM is summarized in Algorithm 1.

V. Experiments

In the experiments, we evaluate the performance of the proposed H-LSTCM compared with the state-of-the-art methods and some baselines on three public datasets.

A. Datasets

Three public datasets used in the experiments are described as follows:

- **BIT dataset [6].** It consists of eight classes of human interactions, i.e., bow, boxing, handshake, high-five, hug, kick, pat, and push. Each class includes 50 videos with cluttered backgrounds. Following in [1], 34 videos per class are randomly chosen as the training data, and the remaining ones are used for testing.
- **UT dataset [29].** It consists of ten videos, where each video contains six classes of human interactions, i.e., handshake, hug, kick, point, punch, and push. After extracting the frames, we obtain 60 video clips, namely 10 video clips per class. Leave-one-out cross-validation is adopted for the experiments.
- **Collective Activity Dataset (CAD) [19].** It contains 44 videos of five multiple-person activities, i.e., crossing, waiting, queuing, walking, and talking. Similar to [20], [30], we select one-third of the video clips from each activity category to form the test set, and the rest of the video clips are used for training. The one-versus-all technique is employed for this recognition task.

TABLE I
RECOGNITION ACCURACY (%) ON THE BIT DATASET.

Method	bow	boxing	handshake	high-five	hug	kick	pat	push	Average
Lan *et al.* [20]	81.25	75.00	81.25	87.50	87.50	81.25	81.25	81.25	82.03
Liu *et al.* [16]	100.00	75.00	81.25	87.50	93.75	87.50	75.00	75.00	84.37
Kong *et al.* [6]	81.25	81.25	81.25	93.75	93.75	81.25	81.25	87.50	85.16
Kong *et al.* [5]	87.50	81.25	87.50	81.25	87.50	81.25	87.50	87.50	85.38
Kong *et al.* [1]	93.75	87.50	93.75	93.75	93.75	87.50	87.50	87.50	90.63
Donahue *et al.* [9]	100.00	75.00	85.00	69.75	85.00	69.75	80.00	76.50	80.13
Ke *et al.* [14]	-	-	-	-	-	-	-	-	85.20
B1	100.00	75.00	62.50	56.25	93.75	68.75	56.25	62.50	71.88
B2	100.00	75.00	84.50	84.50	88.00	88.00	70.00	78.00	83.50
B3	100.00	79.00	84.50	84.50	94.75	88.00	80.50	90.00	87.66
B4	100.00	82.00	85.75	84.50	94.75	88.00	83.00	90.00	88.50
Co-LSTSM [15]	100.00	90.50	92.50	92.50	94.75	88.00	90.50	94.25	92.88
H-LSTCM	100.00	92.50	94.75	95.50	94.75	89.50	91.00	94.25	94.03

B. Implementation Details

In the preprocessing step, the person's bounding box (tracklet) is detected and tracked over all frames by an object detector [31] and an object tracker [32]. Following in [9], the pretrained AlexNet model [33] is employed to extract the fc6 feature (static feature) on each person's bounding box. For the BIT, UT, and CAD datasets, the length T of the time steps is set to 30, 40, and 10, respectively. In the configurations of H-LSTCM on three datasets, the number of memory cell nodes of each Single-Person LSTM, the number of output nodes of each Single-Person LSTM, and the number of sub-memory cell nodes of Co-LSTM are set to 2048, 1024, and 512, respectively. We use the Torch toolbox and Caffe [34] as the deep learning platform and an NVIDIA Tesla K20 GPU to run the experiments. The learning rate, momentum, and decay rate are set to 0.5×10^{-3}, 0.9, and 0.95, respectively. In the experiments, the training of H-LSTCM arrives convergence after approximately 500, 600, and 600 epochs on the BIT, UT, and CAD datasets, respectively.

In the experiments, four baselines are chosen to illustrate the novelty of the proposed H-LSTCM, as follows:

B1 Person-box CNN. The pre-trained AlexNet is deployed on each person's bounding box, where the fc6 features of all persons at one time step are concatenated into a long vector. Then, the concatenated features at all time steps are pooled into a single feature. All features from each video clip are trained and tested by the softmax classifier. This baseline illustrates the importance of deep features.

B2 One CNN + LSTM. This baseline treats two individual actions as a whole. First, multiple person's bounding boxes at a time step are merged into a larger bounding box. Second, the fc6 features are extracted by AlexNet on this "larger" bounding box at each time step. Third, we use the fc6 features as inputs to train an LSTM. This baseline is similar to that of Long-term Recurrent Convolutional Networks [9].

B3 Multiple CNN + LSTM. This baseline models the individual dynamics of multiple persons by Multiple LSTMs. First, AlexNet is deployed on each person's bounding box at each time step to extract the fc6 feature of each person. Second, the extracted fc6 features of each person are fed into an LSTM network to capture the individual dynamic. Third, we average the softmax scores output of all LSTM networks. Here, the averaged score reflects the probability of the activity class. This baseline is similar to that of Two-Stream Convolutional Networks [35].

B4 Single-Person LSTMs + Whole LSTM. This baseline learns the single-person dynamics via multiple LSTMs, and the outputs are pooled into the following LSTM. Specifically, we first use AlexNet to extract fc6 features on person's bounding boxes. Second, the fc6 features of each person are fed into each traditional LSTM network to learn the single-person hidden states. Third, the hidden states of all persons at each time step are max pooled into a single vector, which is fed into the other LSTM network, followed by a softmax. This baseline is similar to that of Hierarchical Deep Temporal Models [13].

C. Results on the BIT dataset

Comparison with baselines. Table I shows the recognition accuracy obtained by the proposed H-LSTCM is better than all baseline methods. Adding the temporal information by employing LSTM (i.e., B2, B3, B4, and Co-LSTSM) improves the performance of B1 without temporal information. Specifically, Co-LSTSM achieves higher accuracy than B2, B3, and B4. It is illustrated that the inter-related motion information among persons is more important than the single-person motion information of individuals for recognizing human interactions. The confusion matrix of H-LSTCM is shown in Appendix C of the supplementary material.

Comparison with state-of-the-art methods. We also compare H-LSTCM with the state-of-the-art methods for human interaction recognition, i.e., hand-crafted spatio-temporal interest points [36] methods of Lan *et al.* [20], Liu *et al.* [16], and Kong *et al.* [1], [5], [6], as well as the LSTM-based methods of Donahue *et al.* [9], and Ke *et al.* [14]. Table I shows the recognition accuracy obtained by all methods, in which some results are reported in [1], [5]. H-LSTCM performs better than the alternatives, especially the LSTM-based methods, i.e., Donahue *et al.* [9] and Ke *et al.* [14]. In particular, H-LSTCM has gained an approximately 9% improvement compared with the state-of-the-art LSTM-based methods (i.e., Ke *et al.* [14] with an accuracy of 85.20%). Some recognition results of H-LSTCM are shown in Figure 3(a).

D. Results on the UT dataset

Comparison with baselines. Table II shows the recognition accuracy of the proposed H-LSTCM compared with that of baselines (including Co-LSTSM [15]). It is observed that H-LSTCM performs better than all the baselines. In particular, H-LSTCM and Co-LSTSM, targeting to model the inter-related dynamics rather than the individual dynamics, achieve impressive accuracy. The confusion matrix of H-LSTCM is shown in Appendix C of the supplementary material.

Comparison with state-of-the-art methods. The proposed H-LSTCM is also compared with the state-of-the-art methods, including some traditional methods (i.e., Ryoo *et al.* [29], Yu *et al.* [37], Kong *et al.* [1], [5], [6], Raptis *et al.* [39], Shariat *et al.* [40], and Zhang *et al.* [7]), a deep learning method (i.e., Wang *et al.* [17]), as well as LSTM-based methods (i.e., Ke *et al.* [14] and Donahue *et al.* [9]). The recognition accuracy results are shown in Table II. The previous Co-LSTSM achieves satisfactory accuracy, i.e., 95%. By further extending Co-LSTSM in a hierarchical way, the proposed

(a) On the BIT dataset
(b) On the UT dataset
(c) On the CAD dataset
(d) On the Volleyball dataset

Fig. 3. Some recognition results of the proposed method on datasets.

TABLE II
RECOGNITION ACCURACY (%) OF DIFFERENT METHODS ON UT DATASET.

Method	handshake	hug	kick	point	punch	push	Average
Ryoo *et al.* [29]	75.00	87.50	62.50	50.00	75.00	75.00	70.80
Yu *et al.* [37]	100.00	65.00	100.00	85.00	75.00	75.00	83.33
Ryoo [38]	80.00	90.00	90.00	80.00	90.00	80.00	85.00
Kong *et al.* [6]	80.00	80.00	100.00	90.00	90.00	90.00	88.33
Kong *et al.* [1]	100.00	90.00	100.00	80.00	90.00	90.00	91.67
Kong *et al.* [5]	90.00	100.00	90.00	100.00	90.00	90.00	93.33
Raptis et al. [39]	100.00	100.00	90.00	100.00	80.00	90.00	93.30
Shariat et al. [40]	-	-	-	-	-	-	91.57
Zhang *et al.* [7]	100.00	100.00	100.00	90.00	90.00	90.00	95.00
Donahue *et al.* [9]	90.00	80.00	90.00	80.00	90.00	80.00	85.00
Ke *et al.* [14]	-	-	-	-	-	-	93.33
Wang *et al.* [17]	-	-	-	-	-	-	95.00
B1	90.00	80.00	80.00	80.00	80.00	80.00	81.67
B2	90.00	80.00	90.00	80.00	90.00	80.00	85.00
B3	100.00	100.00	90.00	80.00	90.00	80.00	90.00
B4	100.00	100.00	90.00	90.00	90.00	80.00	91.67
Co-LSTSM [15]	100.00	100.00	90.00	100.00	90.00	90.00	95.00
H-LSTCM	100.00	100.00	100.00	100.00	100.00	90.00	98.33

H-LSTCM, which first models single-person dynamics and then captures concurrently inter-related dynamics among persons, improves the recognition accuracy to 98.33%, which is the state-of-the-art performance. Some recognition results of H-LSTCM are shown in Figure 3(b).

E. Results on the CAD dataset

Comparison with baselines. We compare the recognition accuracy of the proposed H-LSTCM with that of all the baselines. We also regard the preliminary Co-LSTSM [15] as a baseline. Since most of the group activities in the CAD dataset contain multiple interacting persons ($\geq$ 3 persons), the original Co-LSTSM [15] modeling two interacting persons cannot directly model group activity with multiple persons ($\geq$ 3 persons). Thus, we extend the Co-LSTSM to a new version, named as Co-LSTSM$^+$. Co-LSTSM$^+$ has multiple sub-memory units corresponding to multiple persons, and its architecture is similar to the Co-LSTM module of H-LSTCM in Figure 3. The recognition accuracy of the proposed H-LSTCM and all baselines is shown in Table III. It is observed that H-LSTCM achieves the best performance. We also find that Co-LSTSM+ no longer achieves the significant performance improvements compared with B4. Here, Co-LSTSM$^+$ cannot directly capture the complex inter-related dynamics among multiple persons based on the single-person CNN features, since the collective activities in the CAD dataset are more complex than the interactions in either the BIT dataset or UT dataset. The confusion matrix of H-LSTCM is shown in Appendix C of the supplementary material.

TABLE III
RECOGNITION ACCURACY (%) OF DIFFERENT METHODS ON CAD.

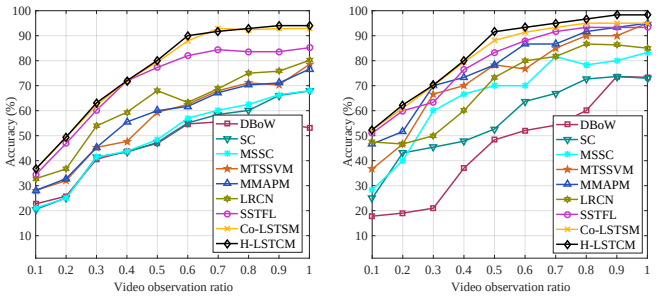
Method	crossing	waiting	queuing	walking	talking	Average
Choi *et al.* [19]	55.4	64.6	63.3	57.9	83.6	65.9
Lan *et al.* [18]	75	74	74	57	61	68.2
Choi *et al.* [3]	76.4	76.4	78.7	36.8	85.7	70.9
Antic *et al.* [41]	73.70	74.50	90.10	62.00	70.00	74.1
Liu *et al.* [16]	72.73	66.67	71.43	83.33	85.71	76.19
Wang *et al.* [4]	64.8	66.0	66.7	89.2	99.5	77.2
Lan *et al.* [20]	68	69	76	80	99	79.7
Choi *et al.* [22]	61.3	82.9	95.4	65.1	94.9	79.9
Kong *et al.* [1]	77.27	77.78	85.71	83.33	100	82.54
Zhou *et al.* [42]	76.83	74.36	93.76	87.63	98.16	82.07
Ibrahim *et al.* [13]	61.54	66.44	96.77	80.41	99.45	81.50
Hajimirsadeghi *et al.* [43]	72	75	92	70	99	81.6
Deng *et al.* [24]	-	-	-	-	-	80.6
Deng *et al.* [25]	-	-	-	-	-	81.2
B1	46.21	53.69	70.20	61.19	74.33	61.12
B2	52.38	54.50	73.89	61.45	76.35	63.71
B3	52.46	54.61	82.00	61.21	79.85	66.02
B4	62.60	65.25	90.74	78.33	95.36	78.46
Co-LSTSM$^+$	65.50	64.85	94.67	75.33	95.33	79.14
H-LSTCM	65.50	68.29	97.90	87.69	99.35	83.75

Comparison with state-of-the-art methods. We also compare the recognition accuracy of H-LSTCM and the state-of-the-art methods, including some traditional methods (i.e., Choi *et al.* [19], Lan *et al.* [18], Choi *et al.* [3], Antic *et al.* [41], Liu *et al.* [16], Wang *et al.* [4], Lan *et al.* [20], Choi *et al.* [22], Kong *et al.* [1], Zhou *et al.* [42], and Hajimirsadeghi *et al.* [43]), a deep learning based method (i.e., Deng *et al.* [24]), RNN based methods (i.e., Deng *et al.* [13], [25], and an LSTM based method (i.e., Ibrahim *et al.* [13]). The recognition accuracy results of all methods are shown in Table II. H-LSTCM achieves better performance than that of the other methods. As a new exploration that leverages the variants of LSTM, the proposed H-LSTCM achieves an approximately 2% improvement compared with the most closely related work [13] (i.e., 81.5% of Ibrahim *et al.* [13]), which uses only the traditional LSTM without any change. Some examples of the recognition results of H-LSTCM are shown in Figure 3(c).

F. Evaluation on Human Interaction Prediction

We also evaluate H-LSTCM on human interaction prediction. In contrast to human interaction recognition, human interaction prediction is defined as recognizing an ongoing interaction activity before the interaction is completely executed [14], [38]. Due to the large variations in the appearance and evolution of scenes, human interaction prediction is a challenging task. Following the experimental setting in [14], [44], a testing video clip is divided into 10 incomplete action executions by using 10 observation ratios (i.e., from 0 to 1 with a step size of 0.1), which represent the increasing amount of sequential data with time. For example, given a testing video clip of length T, an observation ratio of 0.3 denotes that the accuracy is tested with the first $0.3 \times T$ frames. When the observation ratio is 1, namely the entire video clip is used, H-LSTCM acts as a human interaction recognition model.

The baselines include Dynamic Bag-of-Words (DBoW) [38], Sparse Coding (SC) [45], Sparse Coding with Mixture of training video Segments (MSSC) [45], Multiple Temporal Scales based on SVM (MTSSVM) [46], Max-Margin Action Prediction Machine (MMAP-M) [44], Long-term Recurrent Convolutional Networks (LRCN) [9], Spatial-Structural-Temporal Feature Learning (SSTFL) [14] and our



(a) On the BIT dataset. (b) On the UT dataset.

Fig. 4. Comparisons of human interaction prediction on BIT and UT.

preliminary Co-LSTSM [15]. The results of all the methods on the BIT and UT datasets with different observation ratios are listed in Figure 4(a) and Figure 4(b), respectively. Overall, H-LSTCM and Co-LSTSM outperform all the other methods. Here, since all the interactions in the BIT dataset are the two persons interactions with a simple background, the performance of H-LSTCM is comparable to Co-LSTSM on BIT. Specifically, we can observe that: 1) the improvements in H-LSTCM and Co-LSTSM on BIT are more significant when the observation ratio is 0.6; 2) the accuracy of H-LSTCM becomes stable on both BID and UT when the observation ratio is approximately 0.8, which illustrates the close interaction is ending; and 3) since H-LSTSM and Co-LSTSM can accumulate the temporal interacting information, their accuracy monotonously increases with increasing video observation ratio.

VI. Extended Method

A. Extended H-LSTCM

We have validated the effectiveness of the proposed H-LSTCM on human interaction recognition in the single-group activity scene. However, the multiple-person interaction scene sometimes includes multiple sub-groups. For example, in a volleyball game, there are two sub-groups of players from two teams. The players on the same team have more interactions among themselves than with players on the other teams.

To recognize the human interaction within multiple sub-groups, we extend the proposed H-LSTCM to a new version, called Extended H-LSTCM (E-H-LSTCM), as shown in Figure 4. The main extension of E-H-LSTCM is that we use two Co-LSTMs to model the inter-related dynamics of two teams, respectively. Here, the outputs of two Co-LSTMs at one time step are concatenated into a representation of the whole group by a concatenation operation. And then the representations at all time steps are input into the following LSTM layer. Likewise, we extend the baseline Co-LSTSM$^+$ (introduced in Section V-E) to the Extended Co-LSTSM$^+$ (E-Co-LSTSM$^+$) in this way. In the baseline B2, we model the dynamics of each team by B2, and then add a concatenation operation and an LSTM layer on the top of the original LSTM layer, called Extended B2 (E-B2). In the baseline B4, the outputs of multiple Single-Person LSTMs corresponding to one team are pooled into a sequence of representations. The representations of the two teams are concatenated into a long representation, which is then fed into an LSTM layer, called the Extended B4 (E-B4).

We conduct the experiments on the Volleyball Dataset (VD) [13] to evaluate the performance of E-H-LSTCM. This dataset contains 55 volleyball game videos with 4830 annotated frames. In each video, there are two sub-groups in the interaction scene. Each frame provides a group-level activity class label (e.g., left_pass, right_pass, left_set, right_set, left_spike, right_spike, left_winpoint or right_winpoint), and the person's bounding boxes of each person (player). Following

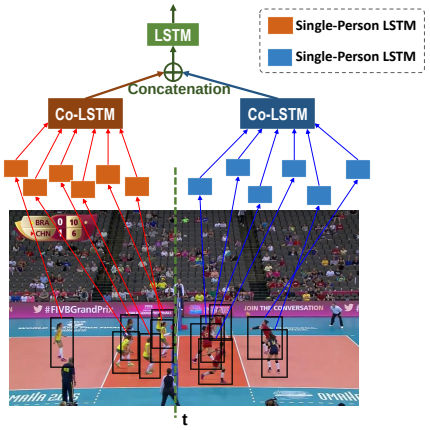


Fig. 5. Framework of E-H-LSTCM (a extensive version of H-LSTCM) on the Volleyball activity with two sub-groups of persons.

TABLE IV
Recognition accuracy (%) on Volleyball dataset.

Method	lpass	rpass	lset	rset	lspike	rspike	lwin	rwin	Average
Ibrahim *et al. [13]*	77.9	81.4	84.5	68.8	89.4	85.6	88.2	87.4	82.9
Shu *et al.* [47]	-	-	-	-	-	-	-	-	83.6
Li *et al. [48]*	55.8	69.1	67.3	52.1	82.1	79.2	-	-	67.6
Biswas *et al. [49]*	-	-	-	-	-	-	-	-	83.0
B1	62.8	62.1	71.4	58.7	65.1	76.5	63.7	61.6	65.2
E-B2	64.6	66.5	76.5	62.7	77.7	74.0	70.6	68.0	70.1
B3	74.4	77.3	81.8	69.7	88.2	83.7	78.6	78.0	79.0
E-B4	77.0	80.9	84.1	68.3	88.8	85.3	88.0	87.7	82.5
E-Co-LSTSM$^+$	81.3	79.5	85.1	70.7	88.8	85.5	88.7	86.9	83.3
E-H-LSTCM	83.9	88.1	90.3	80.4	93.4	89.8	88.7	92.4	88.4

the experimental setting in [13], two-thirds of the annotated frames are used for training, and the remaining ones are used for testing. For the VD, the length T of the time steps is set to 10.

B. Result and Analysis

Comparison with baselines. The recognition accuracy of E-H-LSTCM and all the baselines is shown in Table IV, where "lpass", "rpass", "lset", "rset", "lspike", "rspike", "lwin" and "rwin" denote left_pass, right_pass, left_set, right_set, left_spike, right_spike, left_winpoint and right_winpoint, respectively. E-H-LSTCM achieves the best performance compared with all baseline methods. It is noted that Co-LSTSM+ is comparable to B4, which illustrates that Co-LSTSM+ cannot learn concurrently inter-related representations between multiple persons well when a complex pattern of group activity exists. The confusion matrix of E-H-LSTCM is shown in Appendix C of the supplementary material.

Comparison with state-of-the-art methods. Table IV shows the recognition results of E-H-LSTCM and the state-of-the-art methods, including Ibrahim *et al.* [13], Shu *et al.* [47], Li *et al.* [48], and Biswas *et al.* [49]. E-H-LSTCM achieves higher recognition accuracy than the alternatives. In particular, E-H-LSTCM, with an accuracy of 88.4%, achieves approximately 5% improvement compared with Shu *et al.* [47]. with an accuracy of 83.6%. This demonstrates that E-H-LSTCM is effective in modeling complex human activity with two sub-groups. Some recognition results of E-H-LSTCM are shown in Figure 3(d).

VII. Conclusion and Future Work

On human interaction recognition, we propose a novel Hierarchical Concurrent Long Short-Term Concurrent Memory (H-LSTCM) to learn the dynamic inter-related representation among all persons from the static singe-person features in a hierarchical way. Specifically, for each person, we first feed her/his static single-person features

into a Single-Person LSTM to learn the single-person dynamic. Subsequently, the outputs of all Single-Person LSTM units are fed into a novel Concurrent LSTM (Co-LSTM) unit, which mainly consists of multiple sub-memory units, and a new co-memory cell. In the Co-LSTM unit, each sub-memory unit stores individual motion information. The concurrent LSTM unit selectively integrates and stores the inter-related motion information among multiple interacting persons from multiple sub-memory units via a new co-memory cell. The proposed method is evaluated on several public datasets and yields promising improvements over the state-of-the-art methods.

H-LSTCM assumes that all dynamics of individuals are inter-related in a human interaction scene, and learns the dynamic inter-related representation among multiple persons in a hierarchical way. Actually, in some human interaction scenes, most dynamics of individuals are inter-related, while a few dynamics of individuals are not inter-related. In the future, we will consider exploring H-LSTCM with a hypergraph architecture by regarding each person and the persons' interaction as a node and edge, respectively. By effectively inferring the robust inter-related dynamics of persons, this new architecture aims to further improve the performance in the recognition tasks of the complex human interactions including two persons' interactions, and multiple ($\geq$3) persons' interactions.

REFERENCES

[1] Y. Kong, Y. Jia, and Y. Fu, "Interactive phrases: Semantic descriptions for human interaction recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 9, pp. 1775–1788, 2014.

[2] 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.

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

[4] Z. Wang, S. Liu, J. Zhang, S. Chen, and Q. Guan, "A spatio-temporal crf for human interaction understanding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 8, pp. 1647–1660, 2017.

[5] Y. Kong and Y. Fu, "Close human interaction recognition using patch-aware models," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 167–178, 2016.

[6] Y. Kong, Y. Jia, and Y. Fu, "Learning human interaction by interactive phrases," in *ECCV*, 2012.

[7] Y. Zhang, X. Liu, M. Chang, W. Ge, and T. Chen, "Spatio-temporal phrases for activity recognition," in *ECCV*, 2012.

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

[9] 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 *CVPR*, 2015.

[10] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *CVPR*, 2015.

[11] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *ICCV*, 2015.

[12] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal lstm with trust gates for 3d human action recognition," in *ECCV*, 2016.

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

[14] Q. Ke, M. Bennamoun, S. An, F. Bossaid, and F. Sohel, "Spatial, structural and temporal feature learning for human interaction prediction," *arXiv*, 2016.

[15] 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 *CVPRW*, 2017.

[16] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in *CVPR*, 2011.

[17] X. Wang and Q. Ji, "Hierarchical context modeling for video event recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 9, pp. 1770–1782, 2017.

[18] T. Lan, Y. Wang, G. Mori, and S. N. Robinovitch, "Retrieving actions in group contexts," in *ECCV*, 2010.

[19] W. Choi, K. Shahid, and S. Savarese, "What are they doing?: Collective activity classification using spatio-temporal relationship among people," in *ICCVW*, 2009.

[20] 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.

[21] M. S. Ryoo and J. K. Aggarwal, "Recognition of composite human activities through context-free grammar based representation," in *CVPR*, 2006.

[22] W. Choi and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," in *ECCV*, 2012.

[23] A. Vahdat, B. Gao, M. Ranjbar, and G. Mori, "A discriminative key pose sequence model for recognizing human interactions," in *ICCVW*, 2011.

[24] Z. Deng, M. Zhai, L. Chen, Y. Liu, S. Muralidharan, M. J. Roshtkhari, and G. Mori, "Deep structured models for group activity recognition," in *BMVC*, 2015.

[25] Z. Deng, A. Vahdat, H. Hu, and G. Mori, "Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition," in *CVPR*, 2016.

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

[27] A. Sadeghian, A. Alahi, and S. Savarese, "Tracking the untrackable: Learning to track multiple cues with long-term dependencies," in *ICCV*, 2017.

[28] 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.

[29] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *ICCV*, 2009.

[30] H. Hajimirsadeghi, W. Yan, A. Vahdat, and G. Mori, "Visual recognition by counting instances: A multi-instance cardinality potential kernel," in *CVPRW*, 2015.

[31] R. B. Girshick, "Fast r-cnn," in *ICCV*, 2015.

[32] A. R. Zamir, A. Dehghan, and M. Shah, "Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs," in *ICCV*, 2012.

[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[34] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM MM*, 2014.

[35] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *NIPS*, 2014.

[36] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *VS-PETS*, 2005.

[37] T.-H. Yu, T.-K. Kim, and R. Cipolla, "Real-time action recognition by spatiotemporal semantic and structural forests," in *BMVC*, 2010.

[38] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *ICCV*, 2011.

[39] M. Raptis and L. Sigal, "Poselet key-framing: A model for human activity recognition," in *CVPR*, 2013.

[40] S. Shariat and V. Pavlovic, "A new adaptive segmental matching measure for human activity recognition," in *ICCV*, 2013.

[41] B. Antic and B. Ommer, "Learning latent constituents for recognition of group activities in video," in *ECCV*, 2014.

[42] Z. Zhou, K. Li, X. He, and M. Li, "A generative model for recognizing mixed group activities in still images," in *IJCAI*, 2016.

[43] H. Hajimirsadeghi and G. Mori, "Multi-instance classification by max-margin training of cardinality-based markov networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 9, pp. 1839–1852, 2017.

[44] Y. Kong and Y. Fu, "Max-margin action prediction machine," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1844–1858, 2016.

[45] Y. Cao, D. P. Barrett, A. Barbu, S. Narayanaswamy, H. Yu, A. Michaux, Y. Lin, S. J. Dickinson, J. M. Siskind, and S. Wang, "Recognize human activities from partially observed videos," in *CVPR*, 2013.

[46] Y. Kong, D. Kit, and Y. Fu, "A discriminative model with multiple temporal scales for action prediction," in *ECCV*, 2014.

[47] T. Shu, S. Todorovic, and S.-C. Zhu, "Cern: Confidence-energy recurrent network for group activity recognition," in *CVPR*, 2016.

[48] X. Li and M. C. Chuah, "SBGAR: semantics based group activity recognition," in *ICCV*, 2017.

[49] S. Biswas and J. Gall, "Structural recurrent neural network (srnn) for group activity analysis," in *WACV*, 2018.