Concurrence-Aware Long Short-Term Sub-Memories for Person-Person Action Recognition

Xiangbo Shu†, Jinhui Tang† *, Guo-Jun Qi§, Yan Song†, Zechao Li†, and Liyan Zhang‡

†School of Computer Science and Engineering, Nanjing University of Science and Technology
‡College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics
§Department of Computer Science, University of Central Florida

{shuxb,jinhuitang,songyan,zechao.li}@njust.edu.cn, guojun.qi@ucf.edu, zhangliyan@nuaa.edu.cn

Abstract

Recently, Long Short-Term Memory (LSTM) has become a popular choice to model individual dynamics for single-person action recognition due to its ability of modeling the temporal information in various ranges of dynamic contexts. However, existing RNN models only focus on capturing the temporal dynamics of the person-person interactions by naively combining the activity dynamics of individuals or modeling them as a whole. This neglects the inter-related dynamics of how person-person interactions change over time. To this end, we propose a novel Concurrence-Aware Long Short-Term Sub-Memories (Co-LSTSM) to model the long-term inter-related dynamics between two interacting people on the bounding boxes covering people. Specifically, for each frame, two sub-memory units store individual motion information, while a concurrent LSTM unit selectively integrates and stores inter-related motion information between interacting people from these two sub-memory units via a new co-memory cell. Experimental results on the BIT and UT datasets show the superiority of Co-LSTSM compared with the state-of-the-art methods.

1. Introduction

Person-person interaction (e.g., handshake, hug, etc), as the basic unit in the human activity, is attracting much attention in the computer vision and pattern recognition communities [17, 16, 4, 31]. During a person-person interaction process, there are usually two individual motions from two interacting people respectively, some of which are concurrently inter-related with each other (e.g., two interacting people are stretching out hands in hug interaction). It has been proven that the concurrently inter-related motions between interacting people are discriminative for recognizing the person-person interactions [9, 13]. In most cases of person-person interaction, the concurrently inter-related motions between two interacting people are either 1) quite symmetrically similar to each other (e.g., two interacting people are handshaking); or 2) not quite similar but are strongly interacting to each other (e.g., person A kicks person B, while person B retreats back).

There are mainly two types of solutions for person-person interaction recognition. One solution (e.g., [17, 16, 4, 31]) is to extract the individual motion descriptors (e.g., spatio-temporal interest points [7]) from interacting people, and then predict the class label of an interaction by infer-
ring the coherence between two individual motions. However, this solution regards the person-person interactions as two single-person actions, which ignores some inter-related motion information and brings in some irrelevant individual motion information. The other solution is to extract the motion descriptors on the interactive regions, and then train an interaction recognition model [13]. However, it is hard to locate interactive region before close interacting.

Usually, the difference between person-person interaction (e.g., boxing interaction and pat interaction) is subtle [23, 8, 26], which brings in the challenge to recognize person-person interaction. Recently, due to the powerful ability of capturing the sequential motion information, Recurrent Neural Networks (RNN) [36], especially Long Short-Term Memory (LSTM) [11], has proven successful on human action recognition tasks [8, 6, 22, 12, 17]. To well address the problem of person-person interaction recognition, we aim to explore the long-term inter-related dynamics between two interacting people by leveraging state-of-the-art LSTM model. However, existing LSTM models only modeling human individual dynamics independently do not consider the concurrently inter-related dynamics between interacting people. A naive way is to either 1) merge the individual actions at preprocessing stage [13] (e.g., consider interacting people as a whole); or 2) utilize two LSTM networks to model the individual dynamics of each interacting person respectively, and then fuse the output sequences from two LSTM networks [22]. However, this neglects the inter-related dynamics between interacting people of how person-person interactions can change over time.

To this end, we propose a novel Concurrence-Aware Long Short-Term Sub-Memories (Co-LSTSM) for person-person interaction recognition by modeling the long-term inter-related dynamics between two interacting people on the bounding boxes covering people. It has the ability to aggregate the inter-related memories from individual memories of interacting people over time, as shown in Figure 1. Specifically, we present a novel concurrent LSTM unit consisting of two sub-memory units that store the individual motion information on the bounding box covering people of each video frame. Following these two sub-memory units, a new co-memory cell selectively integrates and stores the memories from two sub-memory units to reveal the concurrently inter-related motion information between interacting people. Overall, two interacting people in each frame are jointly modeled by a concurrent LSTM unit on the bounding boxes covering people, which outputs the concurrently inter-related hidden representations between interacting people rather than the individual hidden representations from individual human. The stacked concurrent LSTM units are recurrent in a time sequence to capture the concurrently inter-related dynamics between two interacting people over time. Extensive experiments on the widely-used benchmarks well show the superior performance of the proposed Co-LSTSM compared with the state-of-the-art methods and several baselines.

Our main contributions in this work are two-fold: (1) We propose a novel Concurrence-Aware Long Short-Term Memories (Co-LSTSM) to effectively address the problem of person-person interaction recognition. (2) To our best knowledge, our work is the first attempt in modeling concurrently long-term inter-related dynamics over time between multiple motion objects by the variants of LSTM.

2. Related Work

2.1. Human Action Recognition

Human activity recognition aims to automatically understand the activities performed by people [31, 32, 23], including group-person interaction recognition (e.g., walking, queuing, etc) [21, 28, 5, 53], person-object interaction recognition (e.g., some people are eating, while the other people are riding a bike) [11, 4], and person-person interaction recognition [12, 12, 21].

For group-person interaction recognition, one solution used in [21, 28] is to exploit the spatial distribution of human activities and present the spatio-temporal descriptors in capturing the spatial distribution of people. The other solution used in [5, 21, 53] is to track all body parts in a video, and then learn the holistic representations to estimate their collective activities. In particular, instead of treating the two problems (i.e., tracking multiple people and estimating their collective activities) separately, Choi et al. [5] presented a unified framework to simultaneously track people and estimate their collective activities. Besides, Lan et al. [21, 21] proposed to recognize the group-person activities by jointly capturing the group activity, the individual human actions, and the interactions among them.

For person-object interaction recognition, there are usually a number of concurrent individual activities (e.g., some people are riding a bike) and group activities (e.g., some people are walking together). To address this challenge, Amer et al. [11] proposed a spatio-temporal AND-OR graph to jointly model the activity parts, person-person spatio-temporal relations, and person-object context, as well as enable multi-target tracking. Subsequently, Amer et al. [2] used a three-layered AND-OR graph to jointly model group activities, individual actions, and participating objects. A key point is that these methods require a multitude of detectors at different levels.

For person-person interaction recognition, some representative works [17, 12, 40] used several interactive phrases as the latent mid-level feature to infer the person-person interaction from the human individual actions. Interactive phrases incorporating rich human knowledge provide an effective way to represent person-person interactions. How-
ever, the difference of some interactions (e.g., boxing and pat) is too stable to be discriminated only by the interactive phrases. Besides, some person-person interactions are complex, which cannot be described well by a certain amount of interactive phrases. Recently, Kong et al. [13] developed a patch-aware latent SVM to recognize the interactions by inferring the closely interactive regions between interacting people. However, it is hard to capture the interactive regions before close interacting. Moreover, Chang et al. [3] proposed to extract features of each interacting person and then learn an interaction matrix between interacting people.

2.2. RNN-based Action Recognition

As neural nets for handling sequential data with variable length, RNN, especially LSTM, has been successfully applied to action recognition [41, 8, 5, 31, 7]. Many RNN-based action recognition methods are embedded the LSTM layer into Convolutional Neural Networks (CNN) [3, 17, 13]. For example, Wu et al. [37] proposed to train three types of CNNs equipped with LSTM to model the spatial, short-term motion and audio clues corresponding to the inputs of video frames, stacked optical flows, and audio spectrogram, respectively. Besides, some skeleton-based action recognition methods utilized RNN to model the long-term contextual information of all skeletons. For example, Du et al. [9] proposed a multilayer RNN framework to feed the five body parts from human skeletons into five subnets. As the number of layers increases, the representations outputs from several subnets are hierarchically fused to the inputs of the higher layers.

Some works aim to design the specific RNN architecture for the different action recognition tasks [12, 41]. For example, in order to capture the co-occurrences of discriminative joints, Zhu et al. [41] added a mixed-norm regularization penalty to the deep LSTM networks. Moreover, the authors proposed an internal dropout technique to operate on the gates, cells, and output responses of the LSTM nodes. To emphasize on the temporal change of motion information between two consecutive frames with the time, Veeriah et al. [53] proposed a Differential RNN architecture equipped with the Derivative States between the LSTM gates. Recently, Shahroudy et al. [51] proposed a Part-aware LSTM that separates the memory cell into the several sub-cells corresponding to the different body parts and explicitly models the dependencies over spatial and temporal domains concurrently. Likewise, Liu et al. [23] also proposed the similar LSTM architecture by pushing the traditional LSTM-based learning into temporal domains and spatial domains simultaneously.

Unlike existing RNN-based action recognition works, we consider the more challenging action recognition scenario within person-person interactions. To capture the interactive motion information rather than the individual motion information, the proposed Co-LSTSM explicitly models the concurrently inter-related dynamics between interacting people. The most related works [13, 12] either combine the individual dynamics of each person or treat the two interacting people as a whole. To our best knowledge, our work is the first time to model the concurrently long-term inter-related dynamics over time between interacting people by the LSTM-based model.

3. Preliminary: RNN for Individual Action

Given an input video clip \( \{ x_t \in \mathbb{R}^n | t = 1, \cdots, T \} \) with the length \( T \), RNN [26] models its dynamics through a sequence of hidden states \( \{ h_t \in \mathbb{R}^m | t = 1, \cdots, T \} \) with \( M \) hidden units, which can be mapped to an output sequence \( \{ z_t \in \mathbb{R}^k | t = 1, \cdots, T \} \) (\( k \) is the number of the classes of actions), i.e.,

\[
h_t = \varphi(W_{hx} \cdot x_t + W_{hh} \cdot h_{t-1} + b_h); \tag{1}
\]

\[
z_t = \varphi(W_{zh} \cdot h_t + b_z), \tag{2}
\]

where \( \varphi(\cdot) \) denotes \( \text{tanh}(\cdot) \), \( W_{hx} \) and \( W_{zh} \) are the weight matrices, and \( b_h \) is the bias vector. Finally, the output \( z_t \) at time step \( t \) can be solved by a softmax function, i.e.,

\[
y_t = \exp(z_{t,i}) / \sum_{j=1}^{\text{num_classes}} \exp(z_{t,j}), \tag{7}
\]

where the \( j \)-th element \( z_{t,j} \) denotes the encoding of the confidence score on the \( j \)-th class action.

Due to the exponential decay in retaining the context information of video frames, Long Short-Term Memory [11], a variant of RNN, provides a solution by allowing the network to learn when to forget previous hidden states and when to update hidden states given new information [8].

Usually, each LSTM unit contains a memory cell (denoted by \( c_t \)) storing the memory of the input sequence up to the time step \( t \). In order to store the memory w.r.t the motion information in the long time, three types of gates (i.e., input gate \( i_t \), forget gate \( f_t \) and output gate \( o_t \)) are incorporated into the LSTM unit to control what information would enter and leave the memory cell over time [11], as follows,

\[
i_t = \sigma(W_{ix} \cdot x_t + W_{ih} \cdot h_{t-1} + b_i); \tag{3}
\]

\[
f_t = \sigma(W_{fx} \cdot x_t + W_{fh} \cdot h_{t-1} + b_f); \tag{4}
\]

\[
o_t = \sigma(W_{ox} \cdot x_t + W_{oh} \cdot h_{t-1} + b_o), \tag{5}
\]

where \( \sigma(\cdot) \) is a sigmoid function; \( W_{xz} \) and \( W_{zh} \) are the weight matrices; \( b_z \) is the bias vector. In addition to three gates, the memory cell \( c_t \) can be expressed as

\[
c_t = f_t \odot c_{t-1} + i_t \odot g_t, \tag{6}
\]

where \( g_t = \varphi(W_{ox} \cdot x_t + W_{oh} \cdot h_{t-1} + b_g) \), and \( \odot \) denotes the element-wise product. Finally, a hidden state \( h_t \) at time step \( t \) can be expressed as

\[
h_t = o_t \odot \varphi(c_t). \tag{7}
\]
4. The Proposed Co-LSTSM

4.1. The Architecture

For person-person interaction recognition, each video frame contain two concurrent individual actions from interacting people, some of which are inter-related with each other. Existing LSTM models targeting to single-person actions cannot handle the person-person interactions well. As mentioned before, we can roughly treat two interacting people as a whole before training the LSTM network. However, this solution will bring in some individual-special motion information. Besides, we can also model the individual dynamics of each person by two LSTM networks respectively, and then naively combine (e.g., concatenate or pool) the output sequences from two LSTM networks into the final representation. However, it is intuitive that this strategy loses some concurrently inter-related motion information between interacting people.

To this end, we propose a Concurrence-Aware Long Short-Term Memories (Co-LSTSM) to capture the concurrently inter-related dynamics between interacting people rather than the individual dynamics of each person. Our key idea is to develop two sub-memory units to store the individual motion information of each person respectively, and a concurrent LSTM unit to selectively integrate and store the concurrently inter-related motion information between interacting people from the individual motion information. Figure 2 illustrates the architecture of a concurrent LSTM unit of the proposed Co-LSTSM. Overall, the concurrent LSTM unit at each time step consists of two specific sub-memory units, two cell gates, a common output gate and a new co-memory cell. Specifically, these two sub-memory units include their respective input gates, forget gate and a new co-memory cell. These two sub-memory units includes the respective input gates (i.e., \( i^a_t \) and \( i^b_t \)), forget gates (i.e., \( f^a_t \) and \( f^b_t \)), sub-memory cells (i.e., \( c^a_t \) and \( c^b_t \)). In particular, two sub-memory cells (i.e., \( c^a_t \) and \( c^b_t \)) are jointly fed into the co-memory cell \( c_t \), followed by the hidden representation \( h_t \).

Two cell gates \( \pi^a_t \) and \( \pi^b_t \) following the sub-memory unit 1 and the sub-memory unit 1 respectively aim to control what memories from two sub-memory units enter and leave at each time step. Unlike the traditional gates, the cell gate \( \pi^a_t \) (\( s \in \{ a, b \} \)) is activated by a nonlinear function of two inputs \( x^a_t \) and \( x^b_t \) and the past hidden state \( h_{t-1} \), i.e.,

\[
\pi^a_t = \sigma(W_{ax}^a \cdot x^a_t + W_{ah}^a \cdot h_{t-1} + b^a), \quad s \in \{ a, b \},
\]

where \( s \in \{ a, b \}, W_{ax}^a \) are the weight matrices, and \( b^a \) is the bias vector. Based on the consistent interactions between two interacting people, these two cell gates \( \pi^a_t \) (\( s \in \{ a, b \} \)) allow more concurrently inter-related motion information between interacting people to enter the co-memory cell \( c_t \) and contribute to one common hidden state. In this work, the co-memory cell \( c_t \) can be expressed as

\[
c_t = \pi^a_t \odot c^a_t + \pi^b_t \odot c^b_t.
\]

In the concurrent LSTM unit, two sub-memory units share a common output gate \( o_t \). The activation of the cell gate \( o_t \) is similar to the activation of the cell gate, i.e.,

\[
o_t = \sigma(W_{ox} \cdot \begin{bmatrix} x^a_t \\ x^b_t \end{bmatrix} + W_{oh} \cdot h_{t-1} + b_o).
\]

Finally, a hidden state \( h_t \) at time step \( t \) can be expressed as

\[
h_t = o_t \odot \varphi(c_t).
\]

Briefly, at time step \( t \), the proposed Co-LSTSM model proceeds in the following order.
• Compute input gates $i_t^s$ and forget gates $f_t^s$ by Eq (8) and Eq (12), respectively;
• Update sub-memory cells $c_t^s$ by Eq (14);
• Compute cell gates $\pi_t^s$ by Eq (17);
• Compute co-memory gate $c_t$ by Eq (13);
• Compute output gate $o_t$ by Eq (15);
• Output $h_t$ by Eq (15).

4.2. Learning Algorithm

We employ a loss function to learn the model parameters of Co-LSTSM by measuring the deviation between the target class $l_t$ and $y_t$ at time step $t$, i.e., $\ell(y_t, l_t) = -\log y_t l_t$. Both types of loss functions can be minimized by Back Propagation Through Time (BPTT) algorithm [3], which unfolds the Co-LSTSM model over several time steps and then runs the back propagation algorithm to train the model. Specifically, LSTM usually uses the truncated BPTT to prevent the back-propagation errors. The idea is that once the back-propagated error leaves the LSTM unit or gates, it will not be allowed to enter the LSTM unit again. Here, we also do not allow the errors to re-enter the concurrent LSTM unit once they leave the co-memory cell.

5. Experiments

5.1. Dataset

We conduct experiments to evaluate the performance of the proposed Co-LSTSM by comparing with the state-of-the-art methods and some baselines on the following two widely-used benchmarks.

BIT dataset [16]. It consists of eight classes of human interactions, i.e., bow, boxing, handshake, high-five, hug, kick, pat, and push. In each class, there are 50 videos, which are captured in real scenarios within the cluttered backgrounds. For some videos, there are partially occluded bodies, moving objects, as well as devise appearances, backgrounds. For some videos, there are partially occluded bodies, moving objects, as well as devise appearances, backgrounds.

UT dataset [29]. It consists of ten videos, each of which contains six classes of human interactions, i.e., handshake, hug, kick, point, punch and push. These videos are captured with different scales and illuminations. The authors provide the interaction labels for each frame. After extracting the frames, we obtain 60 video clips in total, namely 10 video clips per class. The leave-one-out cross validation training strategy is adopted for the experiments, i.e., nine video clips per class are used for training while the remaining one for cross validation. Finally, averaged accuracy on 10 times is reported as the final performance.

5.2. Implementation Details

In the preprocessing step, the bounding box corresponding to each interacting person is detected and tracked over all frames by an object detector [10] and object tracker [39]. Since some works validated that placing the LSTM network on fc6 of CNN performs better than fc7 of CNN [8], we employ the pre-trained AlexNet model [19] to extract the two types of fc6 features on two bounding boxes around two concurrent people, respectively.

For BIT dataset and UT dataset, the length $T$ of time steps is set to 30 and 40, respectively. The sub-memory cell nodes are set 2048 on both BIT and UT. The time steps of each video clip in BIT dataset and UT dataset are set 30 and 40 respectively. We use Torch toolbox and Caffe as the deep learning platform and a NVIDIA Tesla K20 GPU to run the experiments. The learning rate, momentum and decay rate are set $1 \times 10^{-5}$, 0.9 and 0.95, respectively. We plot the learning curve for training Co-LSTSM model on BIT dataset and UT dataset in Figure 3. We can see that the training of Co-LSTSM begins to converge after about 600 and 1300 epochs on the BIT dataset and the UT dataset, respectively.

In experiments, three baselines are conducted to illustrate the novelty of the proposed Co-LSTSM.

• Person-box CNN. The pre-trained AlexNet model is deployed on two bounding boxes around the two concurrent people at each time step respectively, where two fc6 features corresponding to two interacting people are concatenated into a long vector. Then the concatenated features over all time steps are pooled into a single feature. All features from each video clip are trained and tested on the softmax classifier. This baseline can illustrate the importance of deep features.

• One CNN+LSTM. This baseline treats two individual actions as a whole. First, two bounding boxes corresponding two interacting people at each time step are merged into a bigger bounding box. Second, fc6 features are extracted by AlexNet on this “bigger” bounding box at each time step. Third, we use the fc6 features at each time step as inputs to train a LSTM model. The model of this baseline is similar to Long-term Recurrent Convolutional Networks (LRCN) [8].

• Two CNN+LSTM. This baseline models the individual dynamics of two people by two LSTM networks, respectively. First, AlexNet is deployed on the two bounding boxes around two interacting people at each time step to extract fc6 features. Second, fc6 features from two individuals are feed to one LSTM network to capture the individual dynamics, respectively. Third, the softmax scores output from these two LSTM networks are fused. This idea of this baseline is the...
well as LSTM-based methods of Donahue et al. [8] and Ke et al. [13]. In particular, compared with the state-of-the-art LSTM-based methods (i.e., Ke et al. [13] and Donahue et al. [8]), Co-LSTSM has gained about 8% improvement.

5.4. Results on the UT dataset

Comparison with baselines. Table 1 shows the recognition accuracy of the proposed Co-LSTSM compared with the baselines. It is observed that Co-LSTSM performs consistently better than all baselines. “One CNN+LSTM” and “Two CNN+LSTM” considering the temporal information performs better than “Person-box CNN” without temporal information. In particular, “Two CNN+LSTM” achieves the better performance than “One CNN+LSTM”.

Comparison with state-of-the-art methods. Co-LSTSM is also compared with the state-of-the-art methods, including some traditional methods (i.e., Ryoo et al. [29], Yu et al. [38], Kong et al. [16, 17, 14], Raptis & Sigal [26], Shariat & Pavlovic [31], and Zhang et al. [40]), deep learning method (i.e., Wang et al. [55]), as well as LSTM-based methods (i.e., Ke et al. [13] and Donahue et al. [8]). The comparison results are shown in Table 2. We can see that Co-LSTSM also achieves the state-of-the-arts result, i.e., 95.00% by Zhang et al. [31] and Wang et al. [55]. It is noted that Wang et al. [55] adopted deep context features on the event neighborhood, where the size of event neighborhood need be manually defined in the preprocessing step; Zhang et al. proposed a spatio-temporal phrase to capture a certain number of local movements between interacting people, where the number of local movements increases when the interaction becomes complex. As new exploration by leveraging LSTM model, the proposed Co-LSTSM performs better than other LSTM-based methods, i.e., Donahue et al. [8] and Ke et al. [13].

5.5. Evaluation on Human Interaction Prediction

In this work, we also evaluate the proposed Co-LSTSM on human interaction prediction. Unlike person-person interaction recognition, human interaction prediction is defined to recognize an ongoing interaction activity before the interaction is completely executed [13]. Due to the large variations in appearance and the evolution of scenes, interaction prediction at an early stage is a challenging task.

<table>
<thead>
<tr>
<th>Method</th>
<th>bow</th>
<th>boxing</th>
<th>handshake</th>
<th>high-five</th>
<th>hug</th>
<th>kick</th>
<th>pat</th>
<th>push</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lan et al. [21]</td>
<td>81.25</td>
<td>75.00</td>
<td>81.25</td>
<td>87.50</td>
<td>87.50</td>
<td>81.25</td>
<td>81.25</td>
<td>81.25</td>
<td>82.03</td>
</tr>
<tr>
<td>Liu et al. [22]</td>
<td>100.00</td>
<td>75.00</td>
<td>81.25</td>
<td>87.50</td>
<td>93.75</td>
<td>75.00</td>
<td>75.00</td>
<td>84.37</td>
<td></td>
</tr>
<tr>
<td>Kong et al. [10]</td>
<td>81.25</td>
<td>81.25</td>
<td>81.25</td>
<td>93.75</td>
<td>93.75</td>
<td>81.25</td>
<td>81.25</td>
<td>87.50</td>
<td>85.16</td>
</tr>
<tr>
<td>Kong et al. [14]</td>
<td>87.50</td>
<td>81.25</td>
<td>87.50</td>
<td>81.25</td>
<td>87.50</td>
<td>81.25</td>
<td>87.50</td>
<td>85.38</td>
<td></td>
</tr>
<tr>
<td>Kong et al. [17]</td>
<td>93.75</td>
<td>87.50</td>
<td>93.75</td>
<td>87.50</td>
<td>93.75</td>
<td>87.50</td>
<td>87.50</td>
<td>90.63</td>
<td></td>
</tr>
<tr>
<td>Donahue et al. [8]</td>
<td>100.00</td>
<td>75.00</td>
<td>85.00</td>
<td>69.75</td>
<td>85.00</td>
<td>69.75</td>
<td>80.00</td>
<td>76.50</td>
<td>80.13</td>
</tr>
<tr>
<td>Ke et al. [13]</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>85.20</td>
</tr>
</tbody>
</table>

Person-box CNN 100.00 75.00 84.50 84.50 88.00 88.00 70.00 78.00 83.50
One CNN+LSTM 100.00 75.00 84.50 84.50 88.00 88.00 70.00 78.00 83.50
Two CNN+LSTM 100.00 79.00 84.50 84.50 88.00 88.00 70.00 78.00 83.50
Co-LSTSM 100.00 90.50 92.50 92.50 94.75 88.00 90.50 94.25 92.88

Table 1. Recognition accuracy (%) of different methods on the BIT dataset.

Figure 3. Objective loss curve over training epochs.

The same as Two-Stream Convolutional Networks [32].

5.3. Results on the BIT dataset

Comparison with baselines. Table 1 shows the recognition accuracy of the proposed Co-LSTSM compared with the baselines. As shown in this table, Co-LSTSM significantly outperforms the baseline methods. We can see that adding the temporal information by employing LSTM (i.e., “One CNN+LSTM”, and “Two CNN+LSTM”) improves the performance of “Person-box CNN” without temporal information. In particular, “Two CNN+LSTM” achieves the higher accuracy than “One CNN+LSTM”. It is illustrated that an single LSTM model can capture a single motioning object better than multiple motioning objects.

Comparison with state-of-the-art methods. We also compare Co-LSTSM with the state-of-the-art methods for person-person interaction recognition, i.e., hand-crafted spatio-temporal interest points [2] based methods of Lan et al. [21], Liu et al. [22], and Kong et al. [13, 17, 14], as well as LSTM-based methods of Donahue et al. [8] and Ke et al. [13]. Table 1 lists the experimental results, in which some results are reported in [13, 14]. We can see Co-LSTSM performs better than the comparative methods, especially all LSTM-based methods, i.e., Donahue et al. [8] and Ke et al. [13]. In particular, compared with the state-of-the-art LSTM-based methods (i.e., Ke et al. [13] with 85.20%), Co-LSTSM has gained about 8% improvement.
Following experimental setting in [13, 15], 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 with the length \( T \), a prediction accuracy under an observation ratio of \( 0.3 \) denotes that the accuracy is tested with the first length \( 0.3 \times T \) frames. When the observation ratio is 1, namely the entire video clip is used, Co-LSTSM acts as the person-person interaction recognition model.

The comparative methods includes Dynamic Bag-of-Words (DBoW) [27], Sparse Coding (SC) [3], Sparse Coding with Mixture of training video Segments (MSSC) [3], Multiple Temporal Scales based on SVM (MTSSVM) [18], Max-Margin Action Prediction Machine (MMAPM) [15], Long-term Recurrent Convolutional Networks (LRCN) [8], and Spatial-Structural-Temporal Feature Learning (SSTFL) [12]. The comparison results on the BIT dataset with different observation ratios are listed in Figure 4. Overall, Co-LSTSM outperforms all comparative methods for all observation ratios. Specifically, we can see that 1) the improvement of Co-LSTSM is more significant when the observation ratio is 0.6; 2) the accuracy of Co-LSTSM increases rapidly when the observation ratio is 0.5, which illustrates the close interaction is happening; and 3) the accuracy of Co-LSTSM becomes stable when the observation ratio is 0.7, which illustrates the close interaction is ending.

### 6. Conclusions and Future Work

In this work, for person-person interaction recognition, we propose a novel Concurrence-Aware Long Short-Term Sub-Memories (Co-LSTSM) to aggregate the interactive motions between interacting people over time. Specifical-
References


