Global Context-Aware Attention LSTM Networks for 3D Action Recognition

Jun Liu[†], Gang Wang[‡], Ping Hu[†], Ling-Yu Duan[§], Alex C. Kot[†]
[†] School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore
[‡] Alibaba Group, Hangzhou, China
[§] National Engineering Lab for Video Technology, Peking University, Beijing, China
{jliu029, wanggang, phu005, eackot}@ntu.edu.sg, lingyu@pku.edu.cn

Abstract

Long Short-Term Memory (LSTM) networks have shown superior performance in 3D human action recognition due to their power in modeling the dynamics and dependencies in sequential data. Since not all joints are informative for action analysis and the irrelevant joints often bring a lot of noise, we need to pay more attention to the informative ones. However, original LSTM does not have strong attention capability. Hence we propose a new class of LSTM network, Global Context-Aware Attention LSTM (GCA-LSTM), for 3D action recognition, which is able to selectively focus on the informative joints in the action sequence with the assistance of global contextual information. In order to achieve a reliable attention representation for the action sequence, we further propose a recurrent attention mechanism for our GCA-LSTM network, in which the attention performance is improved iteratively. Experiments show that our end-to-end network can reliably focus on the most informative joints in each frame of the skeleton sequence. Moreover, our network yields state-of-the-art performance on three challenging datasets for 3D action recognition.

1. Introduction

Human action recognition is a very important research problem due to its relevance to a wide range of applications. With the advent of depth sensors, such as Microsoft Kinect, Asus Xtion, and Intel RealSense, action recognition using 3D skeleton sequences has attracted a lot of research attention, and lots of advanced approaches have been proposed [33, 14, 1, 72].

Human actions can be represented by a combination of the movements of skeletal joints in 3D space [67, 11]. However, it does not mean all skeletal joints are informative for action analysis. For example, the movements of the hand joints are very informative for the action *clapping*, while the foot joints' movements are not. Different action sequences often have different informative joints, and in the same sequence, the informativeness degree of a joint may also vary over the frames. Therefore, it is beneficial to selectively focus on the informative joints in each frame, and try to ignore the features of the irrelevant ones, since the latter contribute very little for action recognition, and even bring in noise that can corrupt the performance of action recognition [20]. This selectively focusing mechanism is also called as *attention*, which has been demonstrated to be very effective in various areas, such as speech recognition [7], machine

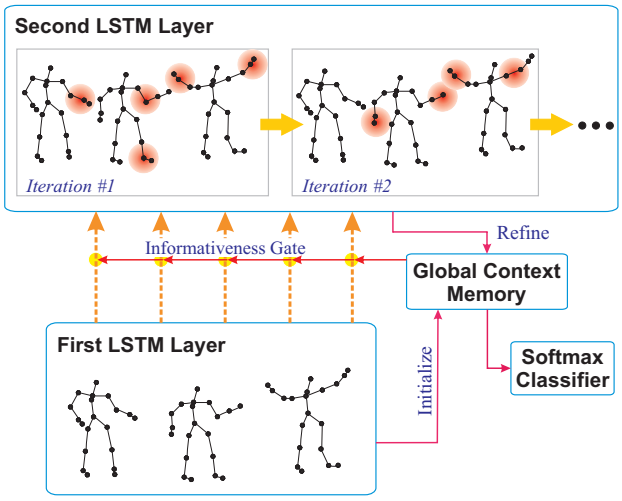

Figure 1. 3D action recognition using the Global Context-Aware Attention LSTM network. The first LSTM layer encodes the skeleton sequence and generates an initial global context memory for this sequence. The second layer performs attention over the inputs with the assistance of global context memory, and further generates an attention representation for the sequence. The attention representation is then used back to refine the global context. Multiple attention iterations are carried out to refine the global context progressively. Finally, the refined global contextual information is used for classification.

translation [3], image caption generation [64], etc.

Recently, Long Short-Term Memory (LSTM) network [15] has been successfully applied to language modeling [46], RGB-based activity analysis [17, 68, 69, 61, 10, 21, 43, 30], and also 3D action recognition [11, 73, 27] due to its strong power in modeling sequential data. However, LSTM does not have strong attention ability for 3D action recognition. This limitation is mainly due to LSTM's restriction in perceiving the global contextual information, which is, however, often important for the global classification problem – 3D action recognition. To perform reliable attention over the joints, we need to measure the informativeness score of each joint in each frame with regarding to the global action sequence. This implies that we need to have global contextual knowledge first. However, the available context at each step of LSTM is relatively local. In LSTM, the sequential data is fed to the network step by step, and the contextual information (hidden representation) of each step is fed to the next one. This indicates at each step, the current available context is the hidden representation from the previous step, which is quite local compared to the global information [1].

In this paper, we extend the original LSTM network and propose Global Context-Aware Attention LSTM (GCA-LSTM) which has strong attention ability for 3D action recognition. In our GCA-LSTM network, the global contextual information is fed to all steps, thus the network can use it to measure the informativeness scores of the new inputs at all steps and accordingly adjust the attention weights for them, i.e., if a new input is informative regarding to the global action, the network imports more information of it, however, if it is irrelevant, the network blocks it.

As shown in Figure 1, our proposed GCA-LSTM network for 3D action recognition contains two LSTM layers. The first layer encodes the skeleton sequence and generates an initial global context memory for it. Then this global context is fed to the second LSTM layer to assist the network to selectively focus on the informative joints in each frame and further produce an attention representation for the global action. Next, the attention representation is fedback to the global context memory to refine it. Specifically, we propose a recurrent attention mechanism for our GCA-LSTM network. Since a refined global context memory is achieved after the attention procedure, we can feed the global context to the second layer again to perform more reliable attention. Multiple attention iterations can be carried out to refine the global context memory progressively. Finally, the refined global context is fed to the classifier to predict the class label of the action.

[1] In LSTM, although the hidden representations of the latter steps contain wider range of contextual information compared to those of the initial steps, their context is still relatively local since LSTM has trouble in remembering information too far in the past [60].

The main contributions of this paper are as follows. (1) We propose a GCA-LSTM network which retains the sequential modeling ability of the original LSTM, meanwhile promoting its selective attention ability. (2) We propose a recurrent attention mechanism to improve the network's attention performance progressively. (3) The visualization results show that the informative joints in each frame of the action sequence can be reliably identified with the assistance of global contextual information. (4) Our end-to-end GCA-LSTM network achieves state-of-the-art performance on all the evaluated datasets.

To the best of our knowledge, this is the first LSTM architecture with explicit attention as its fundamental capability for 3D action recognition.

2. Related Work

3D Action Recognition. Various feature extractors and classifier learning methods for 3D action recognition have been proposed in the past few years [28, 37, 31, 65, 54, 26, 34, 5, 47, 59, 38, 56, 32, 2].

Wang *et al*. [52, 53] proposed an actionlet ensemble model to represent the actions meanwhile capturing the intra-class variances. Vemulapalli *et al*. [49] represented each action as a curve in a Lie group, and adopted a SVM classifier to recognize the actions. Chaudhry *et al*. [4] encoded the skeleton sequences to spatial-temporal hierarchical models, and utilized a set of Linear Dynamical Systems to learn the dynamic structures. Xia *et al*. [62] used Hidden Markov models (HMMs) to model the temporal dynamics in action sequences. Zanfir *et al*. [71] proposed a Moving Pose framework in conjunction with a modified kNN classifier for low-latency activity recognition. Chen *et al*. [6] proposed a part-based 5D feature vector to explore the most relevant joints of body parts in the skeleton sequence. Koniusz *et al*. [22] explored tensor representations to capture the high-order relationships between the skeletal joints. Wang *et al*. [57] introduced a graph-based skeleton motion representation together with a SPGK-kernel SVM for 3D action recognition.

3D Action Recognition Using RNN/LSTM. Besides the aforementioned methods which mainly focus on extracting hand-crafted features, very recently, deep learning, especially recurrent neural network (RNN), based methods have shown their great power in tackling 3D action recognition task. Our proposed network is mainly based on the LSTM network which is an extension of RNN. This part we review the RNN/LSTM based 3D action recognition methods as below since they are very relevant to our approach.

Du *et al*. [11] proposed a hierarchical recurrent neural network to model the human physical structure and temporal dynamics of the skeletal joints. Zhu *et al*. [73] proposed a mixed-norm regularization for the fully connected layers to drive the model to learn co-occurrence features of the

joints. They also introduced an in-depth dropout within the LSTM unit to help train the deep network effectively. Veeriah *et al*. [48] adopted a differential gating mechanism for the LSTM network to make it emphasize on the change of information. Shahroudy *et al*. [35] proposed a Part-aware LSTM network to push the model towards learning the long-term contextual representations for different body parts individually. Liu *et al*. [27] proposed a 2D Spatio-Temporal LSTM framework to employ the hidden sources of action-related information over both spatial and temporal domains concurrently. A trust gate aiming at handling inaccurate 3D coordinates of the skeletal joints was also introduced in [27].

Besides 3D action recognition, RNN and LSTM have also been applied to 3D action detection [25, 18] and forecasting [18].

Unlike the RNN/LSTM based methods mentioned above, which do not consider the informativeness of each joint with regarding to the global action sequence, our GCA-LSTM network performs attention over the evolution steps of LSTM to selectively emphasize on the more informative joints in each frame. An attention representation is generated in our network which can be used to optimize the classification performance. Moreover, a recurrent attention mechanism is introduced to improve the attention performance iteratively.

Attention Mechanism. Our method is also related to the attention mechanism [7, 3, 63, 39, 23, 29, 45] which allows the networks to selectively focus on specific information. Xu *et al*. [64] incorporated soft attention and hard attention for image caption generation. Yao *et al*. [66] introduced a temporal attention mechanism for video caption generation. Luong *et al*. [29] proposed to fuse global attention and local attention for neural machine translation. Stollenga *et al*. [44] proposed a deep attention selective network for image classification.

Although deep learning based methods [40, 36, 55] have been used for action recognition in existing works, most of them do not focus on attention. There are several works which explored attention, such as [39, 58], however, our method is significantly different from them in the following aspects: They all use the state of the previous time step of LSTM, whose contextual information is quite local, to provide the attention scores for the next time step. For global classification problem - action recognition, the global information is necessary for reliably evaluating the importance of each input to achieve reliable attention, thus we propose a global context memory for LSTM, which is used to assess the informativeness score of each input. To the best of our knowledge, we are the first to introduce a global memory cell to LSTM network for global classification problems. Furthermore, we introduce an iterative attention mechanism to promote the attention ability for action recognition, while [39] and [58] use attention only once. Due to our new contributions, our method achieves state-of-the-art performance on all the evaluated datasets.

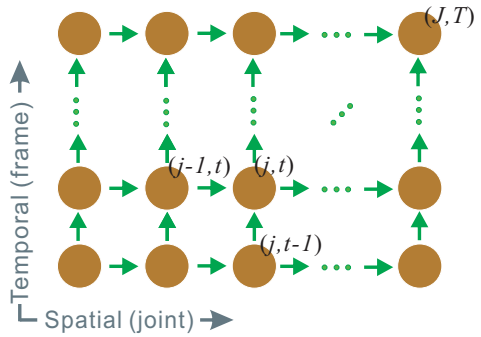

Figure 2. Illustration of the ST-LSTM units [27]. In the spatial direction, the body joints in a frame are arranged as a chain and fed to the network as a sequence. In the temporal direction, body joints are fed over the frames.

3. Global Context-Aware Attention LSTM Networks

In this section, we first briefly review the 2D Spatio-Temporal LSTM (ST-LSTM) as our base network. Then we describe our proposed Global Context-Aware Attention LSTM network in detail, which is capable of selectively focusing on the informative joints in the skeleton sequence with the assistance of global contextual information.

3.1. Spatio-Temporal LSTM

In skeleton-based action recognition, the 3D coordinates of the body joints in each frame are provided. The temporal dependence of the same joint among different frames and the spatial dependence of different joints in the same frame are both important cues for skeleton-based action analysis. Recently, Liu *et al*. [27] proposed a 2D ST-LSTM network for 3D action recognition to model the dependence and contextual information over spatial and temporal domains simultaneously.

In ST-LSTM, the body joints in a frame are arranged and fed as a chain (spatial direction), and the corresponding joints in different frames are also fed in a sequence (temporal direction), as shown in Figure 2. Each ST-LSTM unit is fed with a new input ($x_{j,t}$, 3D location of joint j in frame t), the hidden representation of the same joint at the previous time step ($h_{j,t-1}$), and the hidden representation of the previous joint in the same frame ($h_{j-1,t}$), where $j \in \{1, ..., J\}$ and $t \in \{1, ..., T\}$ denote the indices of joints and frames respectively.

The ST-LSTM unit is equipped with an input gate ($i_{j,t}$), two forget gates corresponding to the two sources of contextual information ($f_{j,t}^{(S)}$ for the spatial domain, and $f_{j,t}^{(T)}$

for the temporal dimension), and an output gate $(o_{j,t})$. The ST-LSTM is formulated as presented in [27]:

$$\begin{pmatrix} i_{j,t} \\ f_{j,t}^{(S)} \\ f_{j,t}^{(T)} \\ o_{j,t} \\ u_{j,t} \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \left(W \begin{pmatrix} x_{j,t} \\ h_{j,t-1} \\ h_{j-1,t} \end{pmatrix} \right) \quad (1)$$

$$\begin{aligned} c_{j,t} &= i_{j,t} \odot u_{j,t} \\ &\quad + f_{j,t}^{(S)} \odot c_{j-1,t} \\ &\quad + f_{j,t}^{(T)} \odot c_{j,t-1} \end{aligned} \quad (2)$$

$$h_{j,t} = o_{j,t} \odot \tanh(c_{j,t}) \quad (3)$$

where $c_{j,t}$ and $h_{j,t}$ denote the cell state and hidden representation of the unit at the spatio-temporal step (j,t), respectively. W is an affine transformation consisting of the model parameters, $u_{j,t}$ is the modulated input, and $\odot$ indicates element-wise product.

3.2. Global Context-Aware Attention LSTM

Previous works [20, 6] have already shown that in each action sequence, there is often a subset of informative joints which are important as they contribute much more to action analysis, while the other ones can be irrelevant (or even noisy) to this action. Consequently, to achieve a high accuracy for 3D action recognition, we need to identify the informative joints and concentrate more on their features, meanwhile trying to ignore the features of the irrelevant ones, i.e., selectively focusing (*attention*) on the informative joints is beneficial for reliable 3D action recognition.

An action can be represented by a combination of the skeletal joints' movements. To reliably identify the informative joints in an action, we can assess the informativeness score of each joint in each frame with regarding to the global action sequence. For this purpose, we need to have global contextual information first. However, the available context at each evolution step of LSTM is the hidden representation from the previous step, which is relatively local compared to the global action. Hence we propose to introduce a global context memory to the LSTM network, which holds the global contextual information for the action sequence and can be fed to each step of LSTM to assist the attention procedure, as shown in Figure 3. We call this LSTM architecture as Global Context-Aware Attention LSTM (GCA-LSTM).

Overview: Our proposed GCA-LSTM network for 3D action recognition is illustrated in Figure 3. It contains three major modules. The *global context memory* maintains an overall representation for the whole action sequence. The *first ST-LSTM layer* encodes the skeleton sequence and initializes the global context memory. The *second ST-LSTM layer* performs attention over the inputs at all spatio-temporal steps to produce an attention representation of the action, which is then used to refine the global context memory. In the first layer, the new input at each spatio-temporal step (j,t) is the 3D coordinates of the joint j in frame t. The inputs of the second layer are the hidden representations from the first layer. Multiple attention iterations (recurrent attention) are carried out in our network to optimize the global context memory iteratively. Finally, the refined global context memory is utilized for classification.

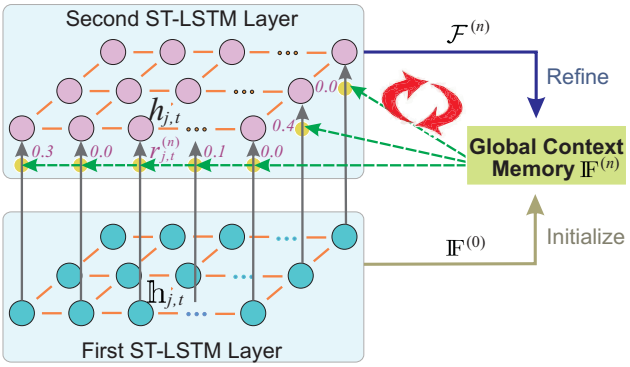

Figure 3. Illustration of the proposed GCA-LSTM network. Some arrows are omitted for clarity.

To facilitate our explanation, in this paper, we use $\mathbb{h}_{j,t}$ instead of $h_{j,t}$ to denote the hidden representation at the step (j,t) in the first layer, and the symbols, such as $h_{j,t}$, $c_{j,t}$, $i_{j,t}$, and $o_{j,t}$, which are defined in Section 3.1, are only used to denote the components in the second layer .

Initializing the Global Context Memory: As our GCA-LSTM network performs attention based on the global contextual information, we need to obtain an initial global context memory first. A feasible scheme is to use the output of the first layer to generate a global context representation. We average the hidden representations from all steps in the first ST-LSTM layer to achieve an initial global context memory as:

$$\mathbb{F}^{(0)} = \frac{1}{JT} \sum_{j=1}^{J} \sum_{t=1}^{T} \mathbb{h}_{j,t} \quad (4)$$

We may also feed all hidden representations of the first layer to a feed-forward neural network, and then use the resultant activation as $\mathbb{F}^{(0)}$. In our experiment, we observe these two initialization choices perform similarly. However, averaging does not involve new parameters, while using a feed-forward network brings considerable amount of parameters.

Attention in the Second ST-LSTM Layer: We assess the informativeness degree of the input at every spatio-temporal step in the second layer. In the n-th attention it-

eration, our network learns an informativeness gate $r_{j,t}^{(n)}$ for each input ($\mathbb{h}_{j,t}$) by feeding the input itself and the global context memory ($\mathbb{F}^{(n-1)}$) produced by the previous attention iteration to a network as:

$$e_{j,t}^{(n)} = W_{e_1}\left(\tanh\left(W_{e_2}\begin{pmatrix}\mathbb{h}_{j,t}\\ \mathbb{F}^{(n-1)}\end{pmatrix}\right)\right) \quad (5)$$

$$r_{j,t}^{(n)} = \frac{\exp(e_{j,t}^{(n)})}{\sum_{p=1}^{J}\sum_{q=1}^{T}\exp(e_{p,q}^{(n)})} \quad (6)$$

where $r_{j,t}^{(n)}$ is the normalized informativeness gate (score) for the input at the step (j,t) in the n-th iteration.

With the learnt informativeness gate $r_{j,t}^{(n)}$, the cell state of the ST-LSTM unit in the second layer can be updated as:

$$\begin{aligned} c_{j,t} &= r_{j,t}^{(n)} \odot i_{j,t} \odot u_{j,t} \\ &\quad + (1 - r_{j,t}^{(n)}) \odot f_{j,t}^{(S)} \odot c_{j-1,t} \\ &\quad + (1 - r_{j,t}^{(n)}) \odot f_{j,t}^{(T)} \odot c_{j,t-1} \end{aligned} \quad (7)$$

This cell state updating scheme can be explained as: if the input ($\mathbb{h}_{j,t}$) is informative (important) regarding to the global context, then we let the learning algorithm update the memory cell of the second ST-LSTM layer by importing more information from it; whereas, if the input is irrelevant, then we need to suppress its effect on the memory and take advantage of more history information.

Refining the Global Context Memory: By adopting the cell state updating scheme in Eq. (7) and then feeding the cell state to Eq. (3), we can obtain the hidden representation $h_{j,t}$ at each step in the second layer, in which joint selection (attention) is involved. The output of the last step in the second layer can be used as an attention representation $\mathcal{F}^{(n)}$ for the action. Finally, the attention representation $\mathcal{F}^{(n)}$ is fed to the global context memory to refine it, as shown in Figure 3. The refinement is formulated as:

$$\mathbb{F}^{(n)} = \mathrm{ReLu}\left(W_F\begin{pmatrix}\mathcal{F}^{(n)}\\ \mathbb{F}^{(n-1)}\end{pmatrix}\right) \quad (8)$$

where $\mathbb{F}^{(n)}$ is the refined version of $\mathbb{F}^{(n-1)}$.

We perform multiple attention iterations (recurrent attention) in our network. The motivation is that after we obtain a refined global context memory, we can carry out the attention again to more reliably identify the informative joints, which can then be used to further refine the global context. After multiple iterations, the global context can be more discriminative for classification.

Learning the Classier: The last refined global context memory $\mathbb{F}^{(N)}$ is fed to a softmax classifier to produce the predicted class label vector $\hat{y}$ as:

$$\hat{y} = \mathrm{softmax}\left(W_c\left(\mathbb{F}^{(N)}\right)\right) \quad (9)$$

The negative log-likelihood loss function [13] is used to measure the difference between the true label y and the predicted result $\hat{y}$. We use the back-propagation through time (BPTT) algorithm to minimize the loss function.

4. Experiments

We validate the proposed approach on the NTU RGB+D dataset [35], UT-Kinect dataset [62], and SBU-Kinect Interaction dataset [70]. To investigate the effectiveness of our network, we conduct extensive experiments with the following three different architectures:

(1) 'ST-LSTM $\oplus$ feed-forward network'. This network structure is similar to the ST-LSTM network in [27]. However, the hidden representations at all spatio-temporal steps of the second layer are concatenated and fed to a one-layer feed-forward network to generate a global representation for the skeleton sequence, and the classification is performed on the *global* representation; while in [27], the classification is performed on single hidden representation at each step (*local* representation), and the prediction scores at all steps are averaged for final classification.

(2) 'GCA-LSTM network'. This is the proposed GCA-LSTM network. The classification is performed on the global context memory.

(3) 'GCA-LSTM network $\ominus$ attention'. This network structure is similar to the above 'GCA-LSTM network', but the attention modules are removed. 'GCA-LSTM network $\ominus$ attention' also has global context representation, which is obtained by averaging the hidden representations at all spatio-temporal steps. Concretely, 'GCA-LSTM network' uses Eq. (7) to update the cell state, while 'GCA-LSTM network $\ominus$ attention' uses the original cell state updating function (Eq. (2)). In 'GCA-LSTM network $\ominus$ attention', the final classification is also performed on the global context representation.

Our experiments are performed based on the Torch7 framework [8]. Stochastic gradient descent (SGD) algorithm is used to train our end-to-end network. We set the learning rate, decay rate, and momentum to 1.5×10^{-3}, 0.95, and 0.9, respectively. The applied dropout probability [42] in our network is 0.5. The dimensions of the cell state of ST-LSTM and the global context memory are both 128. Two attention iterations are performed in our experiment. The first layer is a bi-directional ST-LSTM with trust gates [27]. For a fair comparison, we use the same frame sampling procedure as [27], in which $T = 20$ frames are sampled for each action sequence.

4.1. Experiments on NTU RGB+D Dataset

The NTU RGB+D dataset [35] was recorded with Microsoft Kinect (V2). It contains more than 56 thousand video samples. This dataset includes 60 different action classes. To the best of our knowledge, this is the largest

publicly available dataset for RGB+D based human activity analysis. The large amount of variations in subjects and views make this dataset very challenging.

There are two standard evaluation protocols for this dataset: (1) X-subject: 20 subjects are used for training, and the remaining 20 subjects are for testing; (2) X-view: two view-points are used for training, and one is for testing. To evaluate the proposed approach more extensively, both protocols are tested in our experiment.

We compare our 'GCA-LSTM network' with state-of-the-art methods, as shown in Table 1. We can find that our proposed 'GCA-LSTM network' outperforms the other skeleton-based methods by a large margin. Specifically, the 'GCA-LSTM network' outperforms the 'GCA-LSTM network $\ominus$ attention' and 'ST-LSTM $\oplus$ feed-forward network' on both protocols. This indicates the attention mechanism in our network brings significant performance improvement.

Table 1. Results (accuracies) on the NTU RGB+D dataset.

Method	X-subject	X-view
Skeletal Quads [12]	38.6%	41.4%
Lie Group [49]	50.1%	52.8%
Dynamic Skeletons [16]	60.2%	65.2%
HBRNN [11]	59.1%	64.0%
Deep RNN [35]	56.3%	64.1%
Deep LSTM [35]	60.7%	67.3%
Part-aware LSTM [35]	62.9%	70.3%
ST-LSTM [27]	69.2%	77.7%
'ST-LSTM $\oplus$ feed-forward network'	70.5%	79.5%
'GCA-LSTM network $\ominus$ attention'	70.7%	79.4%
'GCA-LSTM network'	**74.4%**	**82.8%**

As 'ST-LSTM $\oplus$ feed-forward network' and 'GCA-LSTM network $\ominus$ attention' perform classification on the global representations, they both achieve slightly better performance than the original 'ST-LSTM' [27] which performed classification mainly on the local representations. We can also find'ST-LSTM $\oplus$ feed-forward network' and 'GCA-LSTM network $\ominus$ attention' perform similarly. This can be explained as: although their structures seem to be a little different, their fundamental designs are the same. They both use ST-LSTM to model the spatio-temporal dependencies, and perform classification using global information. Moreover, neither of them has explicit attention capability.

Using the NTU RGB+D dataset, we also test the effect of different number of attention iterations on our 'GCA-LSTM network', and show the results in Table 2. We can observe that increasing the iteration number can help to strength the classification performance of our network (using 2 and 3 iterations can obtain higher accuracies compared to using only 1 iteration). However, too many iterations bring performance degradation (the performance of using 3 iterations is slightly worse than that of using 2 iterations). In our experiment, we observe the performance degradation is caused by over-fitting (increasing iteration number introduces new parameters). It is worth noting that the classification results yielded by using the different tested iteration numbers (1, 2, and 3) all outperform the state-of-the-art significantly. We do not try more iterations due to the GPU's memory limitation.

Table 2. Performance (accuracy) comparison for different attention iteration numbers (N) on the NTU RGB+D dataset.

#Iteration	X-subject	X-view
1	71.9%	81.1%
2	74.4%	82.8%
3	72.7%	81.2%

In our method, the informativeness score $r_{j,t}^{(n)}$ is used as a gate within LSTM neuron, as formulated in Eq. (7). We also explore to replace this scheme with soft attention [64, 29], i.e., the attention representation $\mathcal{F}^{(n)}$ is calculated as $\sum_{j=1}^{J}\sum_{t=1}^{T} r_{j,t}^{(n)}\mathbb{h}_{j,t}$. Using the soft attention, the accuracy drops about one percentage point on the NTU RGB+D dataset. This can be explained as equipping LSTM neuron with gate $r_{j,t}^{(n)}$ provides LSTM better insight about when to update, forget or remember. Besides, it can keep the sequential ordering information of the inputs $\mathbb{h}_{j,t}$, while soft attention loses ordering and positional information.

4.2. Experiments on UT-Kinect Dataset

The UT-Kinect dataset [62] was collected with a single stationary Kinect. The skeleton sequences in this dataset are very noisy. A total of 10 action classes were performed by 10 subjects, and every action was performed by the same subject twice.

We follow the standard leave-one-out-cross-validation (LOOCV) protocol in [62] to evaluate our network. Our method achieves state-of-the-art performance on this dataset, as shown in Table 3.

Table 3. Results on the UT-Kinect dataset.

Method	Accuracy
Histogram of 3D Joints [62]	90.9%
Riemannian Manifold [9]	91.5%
Grassmann Manifold [41]	88.5%
Action-Snippets and Activated Simplices [50]	96.5%
Key-Pose-Motifs Mining [51]	93.5%
ST-LSTM [27]	97.0%
'ST-LSTM $\oplus$ feed-forward network'	97.0%
'GCA-LSTM network $\ominus$ attention'	97.5%
'GCA-LSTM network'	**98.5%**

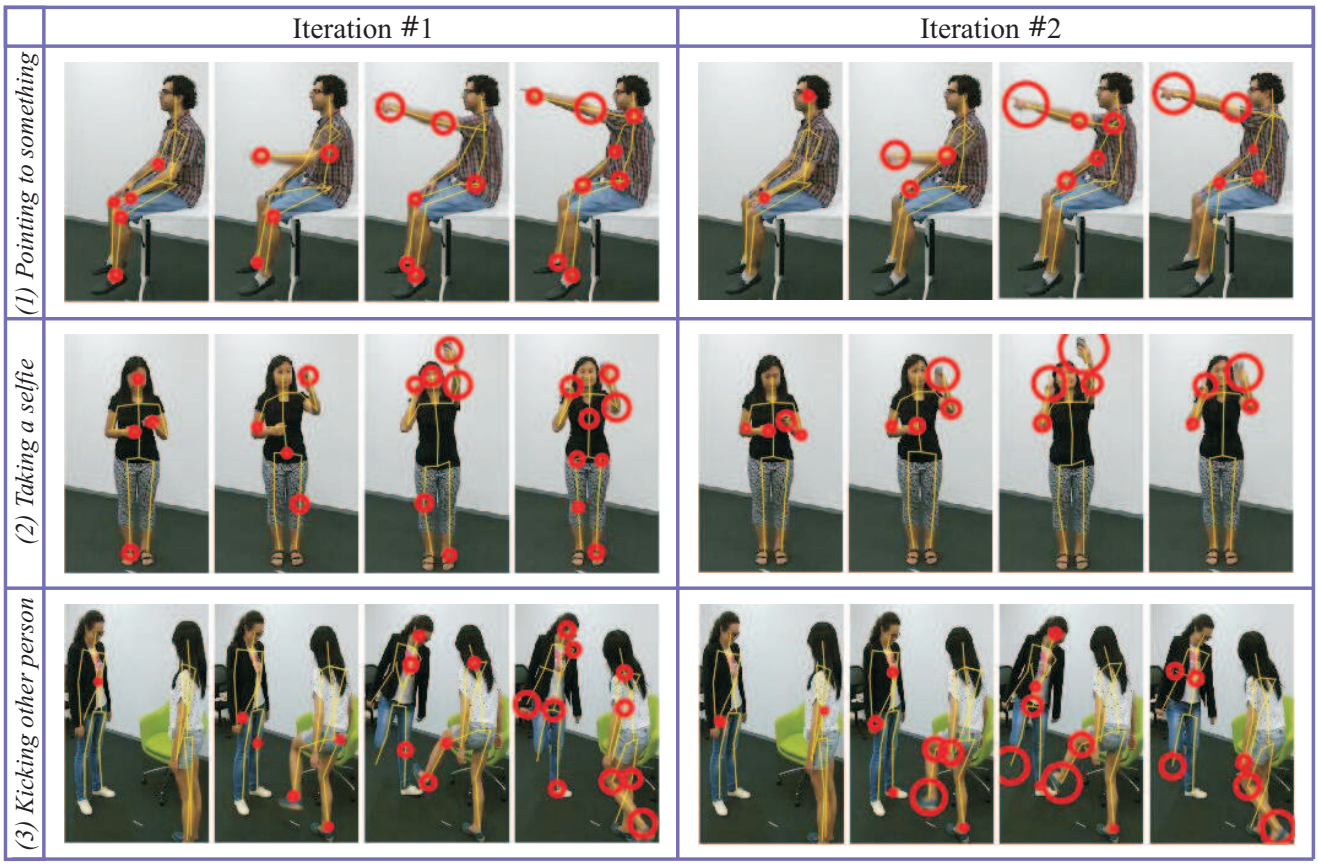

Figure 4. Examples of qualitative results on the NTU RGB+D dataset. Three actions (*pointing to something*, *taking a selfie*, and *kicking other person*) are illustrated. The informativeness gates for two attention iterations are visualized. Four frames are shown for each iteration. The circle size indicates the magnitude of the informativeness gate for the corresponding joint in a frame. For clarity, the joints with tiny informativeness gates are not shown.

4.3. Experiments on SBU-Kinect Interaction Dataset

The SBU-Kinect Interaction dataset [70] contains 8 classes for the purpose of two-person interaction recognition. This dataset includes 282 skeleton sequences corresponding to 6822 frames. This dataset is challenging due to (1) the relatively low accuracy of the joint locations provided by Kinect, and (2) complicated interactions between the two persons in many sequences.

We perform 5-fold cross validation on this dataset by following the standard evaluation protocol in [70]. The experimental results are shown in Table 4. In this table, HBRNN [11], Co-occurrence LSTM [73], Deep LSTM [73], and ST-LSTM [27] are all RNN/LSTM based models for 3D action recognition, and are highly relevant to our method. We can see that our 'GCA-LSTM network' yields the best performance among all of these methods.

4.4. Visualization and Discussion

In order to better understand our network, we analyze and visualize the informativeness score ($r_{j,t}^{(n)}$) learnt by using the global contextual information on the NTU RGB+D dataset in this section.

We analyze the variations of the informativeness scores over the two iterations to verify the effectiveness of the recurrent attention mechanism in our network, and show the

Table 4. Results on the SBU-Kinect Interaction dataset.

Method	Accuracy
Yun *et al*. [70]	80.3%
CHARM [24]	83.9%
Ji *et al*. [19]	86.9%
HBRNN [11]	80.4%
Co-occurrence LSTM [73]	90.4%
Deep LSTM (reported by [73])	86.0%
ST-LSTM [27]	93.3%
'GCA-LSTM network'	**94.1%**

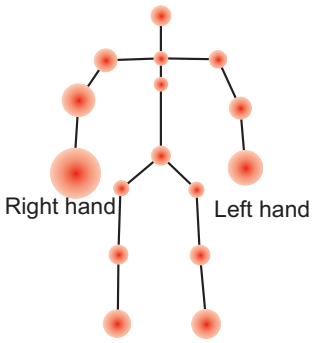

Figure 5. Visualization of the average informativeness gates for all testing samples. The size of the circle around each joint indicates the magnitude of the corresponding informativeness gate.

qualitative results of three actions (*pointing to something*, *taking a selfie*, and *kicking other person*) in Figure 4. The informativeness scores are normalized with soft attention for visualization. In this figure, we can see that the attention performance increases between the two attention iterations. In the first iteration, the network tries to find the potential informative joints over the frames. After this attention, the network achieves a good understanding of the global action. Then in the second iteration, the network can more accurately focus on the informative joints in each frame of the skeleton sequence. We can also find that the informativeness score of the same joint can vary in different frames. This implies *our network performs attention not only in spatial domain, but also in temporal domain*.

To further quantitatively evaluate the effectiveness of the attention mechanism in our network, we analyze the classification accuracies of the three action classes in Figure 4 among all actions. We find if the attention mechanism is not involved, the accuracies of these three classes are 71.7%, 67.7%, and 81.5%, respectively. However, if we use one attention iteration, the accuracies rise to 72.4%, 67.8%, and 83.4%, respectively. If two attention iterations are performed, the accuracies become 73.6%, 67.9%, and 86.6%, respectively.

To roughly explore which joints are more informative for the activities in the NTU RGB+D dataset, we also try to average the informativeness scores for the same joint in all testing sequences, and visualize it in Figure 5. We can find that averagely, more attention is assigned to the hand and foot joints. This is because in the NTU RGB+D dataset, most of the actions are related to the hand and foot postures and motions. We can also observe that the average informativeness score of the right hand joint is higher than that of left hand joint. This indicates most of the subjects are right-handed.

5. Conclusion

In this paper, we extend the LSTM network to achieve a Global Context-Aware Attention LSTM (GCA-LSTM) network for 3D action recognition, which has strong capability in selectively focusing on the informative joints in each frame of the skeleton sequence with the assistance of global contextual information. We further propose a recurrent attention mechanism for our GCA-LSTM network, in which the selectively focusing ability is strengthened iteratively. The experimental results validate the contributions by achieving state-of-the-art performance on all the evaluated benchmark datasets.

Acknowledgement

This research was carried out at the Rapid-Rich Object Search (ROSE) Lab at Nanyang Technological University (NTU), Singapore.

The ROSE Lab is supported by the National Research Foundation, Singapore, under its Interactive Digital Media (IDM) Strategic Research Programme.

The research is in part supported by Singapore Ministry of Education (MOE) Tier 2 ARC28/14, and Singapore A*STAR Science and Engineering Research Council PSF1321202099.

We gratefully acknowledge the support of NVAITC (NVIDIA AI Technology Centre) for the donation of Tesla K40 and K80 GPUs used for our research at the ROSE Lab. Jun Liu would like to thank Kamila Abdiyeva, Amir Shahroudy and Bing Shuai from NTU, and Peiru Zhu from Alibaba for helpful discussions.

References

[1] J. K. Aggarwal and L. Xia. Human activity recognition from 3d data: A review. *PR Letters*, 2014.

[2] R. Anirudh, P. Turaga, J. Su, and A. Srivastava. Elastic functional coding of human actions: from vector-fields to latent variables. In *CVPR*, 2015.

[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[4] R. Chaudhry, F. Ofli, G. Kurillo, R. Bajcsy, and R. Vidal. Bio-inspired dynamic 3d discriminative skeletal features for human action recognition. In *CVPRW*, 2013.

[5] C. Chen, R. Jafari, and N. Kehtarnavaz. Fusion of depth, skeleton, and inertial data for human action recognition. In *ICASSP*, 2016.

[6] H. Chen, G. Wang, J.-H. Xue, and L. He. A novel hierarchical framework for human action recognition. *PR*, 2016.

[7] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio. Attention-based models for speech recognition. In *NIPS*, 2015.

[8] R. Collobert, K. Kavukcuoglu, and C. Farabet. Torch7: A matlab-like environment for machine learning. In *NIPSW*, 2011.

[9] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo. 3-d human action recognition by shape analysis of motion trajectories on riemannian manifold. *IEEE Transactions on Cybernetics*, 2015.

[10] 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.

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

[12] G. Evangelidis, G. Singh, and R. Horaud. Skeletal quads: Human action recognition using joint quadruples. In *ICPR*, 2014.

[13] A. Graves. Supervised sequence labelling. In *Supervised Sequence Labelling with Recurrent Neural Networks*. 2012.

[14] F. Han, B. Reily, W. Hoff, and H. Zhang. Space-time representation of people based on 3d skeletal data: a review. *arXiv*, 2016.

[15] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 1997.

[16] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. In *CVPR*, 2015.

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

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

[19] Y. Ji, G. Ye, and H. Cheng. Interactive body part contrast mining for human interaction recognition. In *ICMEW*, 2014.

[20] M. Jiang, J. Kong, G. Bebis, and H. Huo. Informative joints based human action recognition using skeleton contexts. *Signal Processing: Image Communication*, 2015.

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

[22] P. Koniusz, A. Cherian, and F. Porikli. Tensor representations via kernel linearization for action recognition from 3d skeletons. In *ECCV*, 2016.

[23] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, and R. Socher. Ask me anything: Dynamic memory networks for natural language processing. In *ICML*, 2016.

[24] W. Li, L. Wen, M. Choo Chuah, and S. Lyu. Category-blind human action recognition: A practical recognition system. In *ICCV*, 2015.

[25] Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan, and J. Liu. Online human action detection using joint classification-regression recurrent neural networks. In *ECCV*, 2016.

[26] I. Lillo, J. Carlos Niebles, and A. Soto. A hierarchical pose-based approach to complex action understanding using dictionaries of actionlets and motion poselets. In *CVPR*, 2016.

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

[28] J. Luo, W. Wang, and H. Qi. Group sparsity and geometry constrained dictionary learning for action recognition from depth maps. In *ICCV*, 2013.

[29] M.-T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, 2015.

[30] S. Ma, L. Sigal, and S. Sclaroff. Learning activity progression in lstms for activity detection and early detection. In *CVPR*, 2016.

[31] M. Meng, H. Drira, M. Daoudi, and J. Boonaert. Human-object interaction recognition by learning the distances between the object and the skeleton joints. In *FG*, 2015.

[32] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. *JVCIR*, 2014.

[33] L. L. Presti and M. La Cascia. 3d skeleton-based human action classification: A survey. *PR*, 2016.

[34] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian. Real time action recognition using histograms of depth gradients and random decision forests. In *WACV*, 2014.

[35] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *CVPR*, 2016.

[36] A. Shahroudy, T.-T. Ng, Y. Gong, and G. Wang. Deep multimodal feature analysis for action recognition in rgb+d videos. *TPAMI*, 2017.

[37] A. Shahroudy, T.-T. Ng, Q. Yang, and G. Wang. Multimodal multipart learning for action recognition in depth videos. *TPAMI*, 2016.

[38] A. Shahroudy, G. Wang, and T.-T. Ng. Multi-modal feature fusion for action recognition in rgb-d sequences. In *ISCCSP*, 2014.

[39] S. Sharma, R. Kiros, and R. Salakhutdinov. Action recognition using visual attention. In *ICLRW*, 2016.

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

[41] R. Slama, H. Wannous, M. Daoudi, and A. Srivastava. Accurate 3d action recognition using learning on the grassmann manifold. *PR*, 2015.

[42] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 2014.

[43] N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised learning of video representations using lstms. In *ICML*, 2015.

[44] M. F. Stollenga, J. Masci, F. Gomez, and J. Schmidhuber. Deep networks with internal selective attention through feedback connections. In *NIPS*, 2014.

[45] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus. End-to-end memory networks. In *NIPS*, 2015.

[46] M. Sundermeyer, R. Schlüter, and H. Ney. Lstm neural networks for language modeling. In *INTERSPEECH*, 2012.

[47] L. Tao and R. Vidal. Moving poselets: A discriminative and interpretable skeletal motion representation for action recognition. In *ICCVW*, 2015.

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

[49] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In *CVPR*, 2014.

[50] C. Wang, J. Flynn, Y. Wang, and A. L. Yuille. Recognizing actions in 3d using action-snippets and activated simplices. In *AAAI*, 2016.

[51] C. Wang, Y. Wang, and A. L. Yuille. Mining 3d key-pose-motifs for action recognition. In *CVPR*, 2016.

[52] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, 2012.

[53] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Learning actionlet ensemble for 3d human action recognition. *TPAMI*, 2014.

[54] J. Wang and Y. Wu. Learning maximum margin temporal warping for action recognition. In *ICCV*, 2013.

[55] P. Wang, W. Li, Z. Gao, Y. Zhang, C. Tang, and P. Ogunbona. Scene flow to action map: A new representation for rgb-d based action recognition with convolutional neural networks. In *CVPR*, 2017.

[56] P. Wang, W. Li, P. Ogunbona, Z. Gao, and H. Zhang. Mining mid-level features for action recognition based on effective skeleton representation. In *DICTA*, 2014.

[57] P. Wang, C. Yuan, W. Hu, B. Li, and Y. Zhang. Graph based skeleton motion representation and similarity measurement for action recognition. In *ECCV*, 2016.

[58] Y. Wang, S. Wang, J. Tang, N. O'Hare, Y. Chang, and B. Li. Hierarchical attention network for action recognition in videos. *arXiv*, 2016.

[59] J. Weng, C. Weng, and J. Yuan. Spatio-temporal naive-bayes nearest-neighbor for skeleton-based action recognition. In *CVPR*, 2017.

[60] J. Weston, S. Chopra, and A. Bordes. Memory networks. In *ICLR*, 2015.

[61] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *ACM MM*, 2015.

[62] L. Xia, C.-C. Chen, and J. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *CVPRW*, 2012.

[63] C. Xiong, S. Merity, and R. Socher. Dynamic memory networks for visual and textual question answering. In *ICML*, 2016.

[64] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[65] X. Yang and Y. Tian. Effective 3d action recognition using eigenjoints. *JVCIR*, 2014.

[66] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *ICCV*, 2015.

[67] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall. A survey on human motion analysis from depth data. In *Time-of-flight and depth imaging. sensors, algorithms, and applications*. 2013.

[68] S. Yeung, O. Russakovsky, G. Mori, and L. Fei-Fei. End-to-end learning of action detection from frame glimpses in videos. In *CVPR*, 2016.

[69] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *CVPR*, 2015.

[70] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In *CVPRW*, 2012.

[71] M. Zanfir, M. Leordeanu, and C. Sminchisescu. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *ICCV*, 2013.

[72] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang. Rgb-d-based action recognition datasets: A survey. *PR*, 2016.

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