Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition

Recent attempts on 3D action recognition suggested to develop RNN-based learning methods to model the contextual dependency in the temporal domain. This paper extends this idea to spatiotemporal domains to analyze the hidden sources of action-related information within the input data over both domains concurrently. Inspired by the graphical structure of the human skeleton, the authors further propose a more powerful tree-structure based traversal method. To handle the noise and occlusion in 3D skeleton data, they introduce new gating mechanism within LSTM to learn the reliability of the sequential input data and accordingly adjust its eﬀect on updating the long-term context information stored in the memory cell.

The main structure of this paper is as follows:

At first, they introduce several recent feature extraction and classier learning approaches for 3D action recognition as well as the merits and demerits of RNN-based methods. They also mention their main contribution: a spatio-temporal long short-term memory (ST-LSTM) with a skeleton-based tree traversal technique and the trust gate.

Then they review recent RNN-based and LSTM-based approaches, all of which concatenate the joint-based input features. In contrast, this paper explicitly models the dependencies between the joints and applies recurrent analysis over spatial and temporal domains concurrently. Besides, it develops a novel trust gate to make LSTM robust to noisy input data.

In the following section, they present details about the ST-LSTM: first, the standard LSTM networks with its transition equations, then the proposed spatio-temporal LSTM model and the skeleton-based tree traversal, both with relevant diagram or equation. Next, an effective gating scheme for LSTM to deal with the measurement noise in the input data (body joint locations) for the task of 3D human action recognition.

Afterwards, they evaluate the performance of three different configurations: ST-LSTM (Joint Chain), ST-LSTM (Tree Traversal), and ST-LSTM (Tree Traversal) + Trust Gate with some other methods on four datasets: NTU RGB+D dataset, SBU Interaction dataset, UT-Kinect dataset, and Berkeley MHAD dataset. They also specifically evaluate noisy samples from MSR Action3D dataset to study the effectiveness of the trust gate in the proposed network model.

Finally, according to the convincing accuracy achieved in the experiments, they conclude the effectiveness of their novel method which surpasses the existing state-of-the-art methods on four evaluated datasets.

There are three main unique properties:

First, Spatio-Temporal LSTM. As shown in Fig. 1, every ST-LSTM unit corresponds to one of the skeletal joints. Each of the units receives the hidden representation of the previous joint and also the hidden representation of its own joint from the previous frame. In this section we assume joints are arranged in a chain-like sequence with the order shown in Fig. 2(a). Each ST-LSTM unit is fed with its input (location of the corresponding joint at current frame), its own hidden representation at the previous time step, and the hidden representation of the previous joint at current frame. Each unit is also equipped with two different forget gates corresponding to the two incoming channels of context information for the spatial domain, and for the temporal domain.

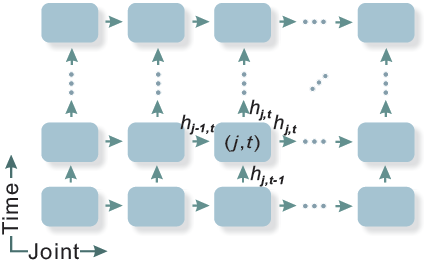


Fig. 1. The illustration of the proposed spatio-temporal LSTM network. In the spatial direction, body joints in a frame are fed in a sequence. In the temporal direction, the locations of the corresponding joints are fed over time. Each unit receives the hidden representation of previous joints and previous frames of the same joint as contextual information.

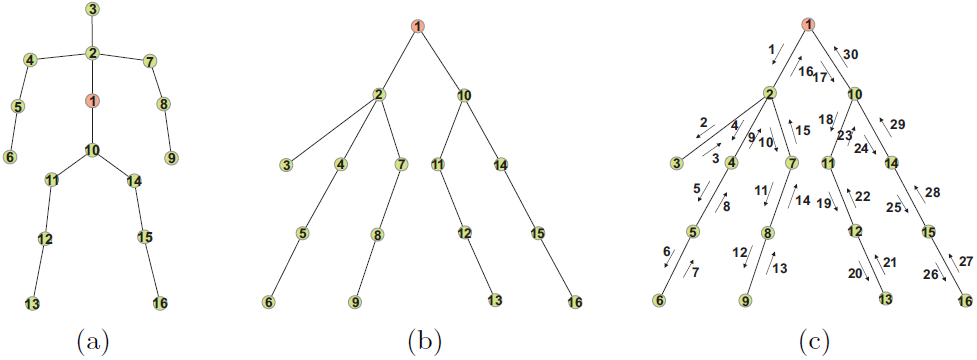


Fig. 2. (a) Skeletal joints of a human body. In the simple joint chain model, the joint visiting order is 1-2-3-...-16. (b) Skeleton is transformed to a tree structure. (c) Tree traversal over the spatial steps. The tree can be unfolded to a chain with the traversal, and the joint visiting order is 1-2-3-2-4-5-6-5-4-2-7-8-9-8-7-2-1-10-11-12-13-12-11-10-14-15-16-15-14-10-1.

Second, a bidirectional tree traversal method to visit joints in a sequence which maintains the adjacency information of the skeletal tree structure. This traversal strategy guarantees the transmission of the data in both directions (top-down and bottom-up) inside the adjacency tree structure. Therefore, each node will have the contextual information from both its descendants and ancestors.

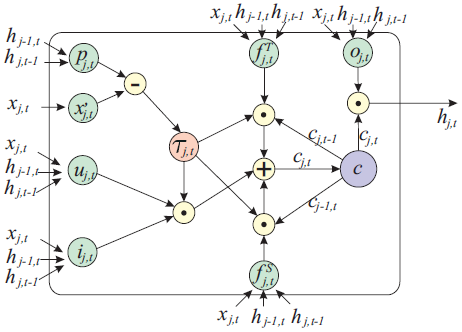


Fig. 3. Schema of the proposed ST-LSTM with trust gate.

Third, a new gate added to the LSTM unit which analyzes the reliability of the input at each spatio-temporal step, based on the estimation of the input from the available contextual information. The amount of the estimation error is used as input to a new “trust gate”. The derived trust value provides information to the long-term memory mechanism to learn better decisions about when and how to remember and forget the contents of the memory cell. For example, when the trust gate finds out the current joint has wrong 3D measurements, it can block the input gate and prevent the memory cell from updating based on current unreliable input. If the new input cannot be trusted (because of noise or occlusion), then it needs to take advantage of more history information and try to block the new input. In contrast, if the input is reliable, it can let the learning algorithm update the memory cell by importing input information.

The first feature can apply in those occasions where the state of a series of objects is determined not only by their own previous state but also interactions among them, like the city weather forecast.

The second feature can be adopted to improve the body coordination of dance robot or even Mars rovers.

The last feature can be helpful to heritage restoration since cultural relics such as inscriptions are usually with minimal damages while the context is inferable.