View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition from Skeleton Data: Overview and comments

Rémy Gaudré, Taha Zoulagh, Sergio Duarte, Moad Hani
MICS
University of Luxembourg
Belval, Luxembourg
remy.gaudre.001@student.uni.lu, taha.zoulagh.001@student.uni.lu,
sergio.duarte.001@student.uni.lu, moad.hani.001@student.uni.lu

***Abstract*—This document is summarising the work presented in View Adaptive Recurrent Neural Networks for High Performance Human Action Recognition from Skeleton Data. This main objective of this work is to propose a novel view adaptation scheme to automatically regulate observation viewpoints during the occurrence of an action.**

***Index Terms*—Adaptive Recurrent Neural Networks, 3D skeleton, Computer Vision, Human Action Recognition**

I. Introduction

Recognition of human behavior can be used in many domains, such as surveillance, interaction between humans and computers, video indexing/retrieval or video comprehension [1]. There already exist some solutions such as Microsoft kinect [2] or Intel RealSense that make 3D skeleton data easily obtainable. The key challenge faced by this identification of behavior is the dynamic differences in perspective due to camera orientation or human orientation. Indeed, the ability of recognizing these actions rely on the location of the skeleton in space. However, model recognition ignores this problem by centering the coordinates of the system on the body. The problem is that there is a lack of information in this processing. For example: the action of walking becomes walking in the same place and the action of dancing with body rotating becomes dancing with body facing a fixed orientation. So to solve this problem, the scientists designed a view adaptive Recurrent Neural Netword with LSTM architecture to learn and determine the appropriate viewpoints based on the input skeleton [3].

So, the main contributions on this step are the following:

- A self-regulated view adaptation scheme is proposed, which aims to re-position the observation viewpoints dynamically to facilitate better recognition of the action from skeleton data.
- The integration of the proposed method into an end-to-end LSTM network which automatically determines the "best" observation viewpoints during recognition.
- Observations and analysis have been made from the view adaptation model. To conclude, the proposed model automatically regulates the skeletons to more consistent observation viewpoints while tracking an action.

II. RNN and LSTM

Traditional neural networks can't do this, and it seems like a major shortcoming. For example, imagine you want to classify what kind of event is happening at every point in a movie. It's unclear how a traditional neural network could use its reasoning about previous events in the film to inform later ones.

Recurrent neural networks address this issue. They are networks with loops in them, allowing information to persist.

A loop allows information to be passed from one step of the

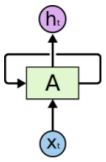

Fig. 1. Neural Networks have loops.

network to the next [4].

Long Short Term Memory networks (LSTM), are a special kind of RNN, capable of learning long-term dependencies.

LSTMs are explicitly designed to avoid the long-term dependency problem. Remembering information for long periods of time is practically their default behavior, not something they struggle to learn [5].

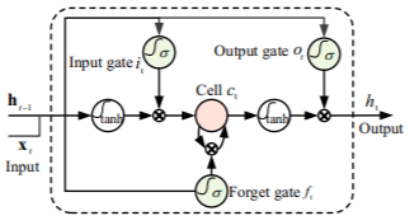

Fig. 2. LSTM.

The LSTM does have the ability to remove or add information to the cell state, carefully regulated by structures called gates.

III. The Self-Regulated View Adaptation Scheme

A. Basic Idea

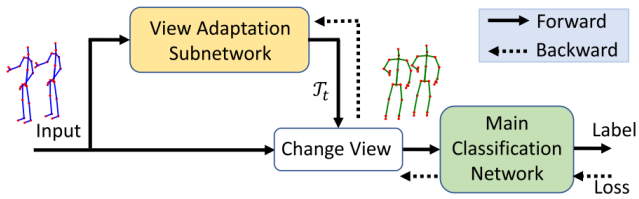

Fig. 3. Flowchart of the end-to-end view adaptive neural network.

We see that the view adaptation Subnetwork acts only on the view. It will change the view of the skeleton input to help the Main Classification Network to determine the class of the action performed by the skeleton. The entire network is end-to-end trained to optimize the classification performance.

B. Problem formulation

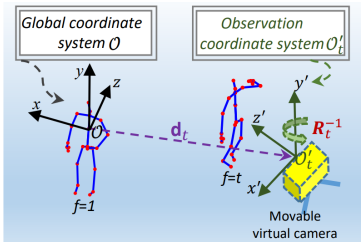

Fig. 4. Schema of the problem formulation

Given a skeleton sequence S with T frames, under the global coordinate system O, the j^{th} skeleton joint on the t^{th} frame is denoted as $v_{t,j} = [x_{t,j}, y_{t,j}, z_{t,j}]^T$, where $t \in (1, ..., T)$, $j \in (1, ..., J)$, J denotes the total number of skeleton joints in a frame.

For the t^{th} frame, assume the movable virtual camera is placed at a suitable viewpoint, with the corresponding observation coordinate system obtained from a translation by $d_t \in R^3$, and a rotation of α_t, β_t, γ_t radians anticlockwise around the X-axis, Y -axis, and Z-axis, respectively, of the global coordinate system. The representation of the j^{th} skeleton under observation coordinate system O'_t is:

$$V'_{t,j} = [x'_{t,j}, y'_{t,j}, z'_{t,j}]^T = R_t * (V'_{t,j} - d_t) \quad (1)$$

R_t can be represented as:

$$R_t = R^x_{t,\alpha} * R^y_{t,\beta} * R^z_{t,\gamma} \quad (2)$$

where $R^x_{t,\alpha}$, $R^y_{t,\beta}$ and $R^z_{t,\gamma}$ denote he coordinate transform for rotating the original coordinate system around X-axis by α_t radians, Y-axis by β_t radians, and Z-axis by γ_t radians anticlockwise, respectively. For instance, we can define:

$$R^y_{t,\beta} = \begin{bmatrix} cos(\beta_t) & sin(\beta_t) & 0 \\ sin(\beta_t) & cos(\beta_t) & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (3)$$

IV. The Integrated Solution

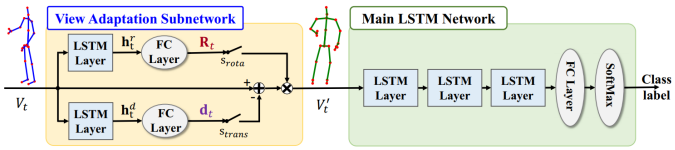

Fig. 5. Architecture of the proposed view adaptive neural networks

The use of a View Adaptation Subnetwork detemines automatically the rotation parameters : α_t, β_t, γ_t to obtain the rotation matrix R_t and and the translation vector d_t.

The main LSTM Network is composed by three successive LSTM layers, followed by one full conncetion (FC) layer with a SoftMax classifier.

End-to-End Training:The number of neurons of the FC layer is equal to the number of action classes. They use cross-entropy loss as the training loss.

V. Observations of Results

The used dataset are the following:

- NTU RGB+D Dataset (NTU) [6]
- SBU Kinect Interaction Dataset (SBU) [7]
- SYSU 3D Human-Object Interaction Set (SYSU) [8]

The relevant obtained results are listed in the tables I, II and III:

Experiment results demonstrate that the proposed model can

Methods	CS	CV
ST-LSTM (Tree Traversal) + Trust Gate [5]	69,2	77,7
STA-LSTM [9]	73,4	81,2
VA-LSTM	79,4	87,6

TABLE I
COMPARISONS ON THE NTU DATASET WITH CROSS-SUBJECT AND CROSS-VIEW SETTINGS IN ACCURACY %

Methods	Acc. (%)
STA-LSTM [9]	91,5
ST-LSTM [5]	93,4
VA-LSTM	97,2

TABLE II
COMPARISONS ON THE SYSU DATASET IN ACCURACY %

Methods	setting-1	setting-2
LAFF [10]	-	54,2
Dynamic Skeleton [8]	75,5	76,9
VA-LSTM	76,9	77,5

TABLE III
COMPARISONS ON THE SBU DATASET IN ACCURACY %

significantly improve the recognition performance on three benchmark datasets and achieve state-of-the-art results.

VI. Conclusion and discussions

The end-to-end view adaptation model for human action recognition from skeleton data was presented. The proposed network is capable of regulating the observation viewpoints to the suitable ones by itself, with the optimization target of maximizing recognition performance. It overcomes the limitations of the human defined pre-processing approaches by exploiting the optimal viewpoints through the content dependent recurrent neuron network model.

On the other hand, the problem of Human Action Recognition from Skeleton Data still remain much intention and is tackled by researchers to improve it and and achieve better performance. So, we can say that this method showed weaknesses against some new approaches on some specific Dataset such that the NTU-RGBD and UT-Kinect datasets. For instance, the obtained accuracy based on Action Recognition with Directed Graph Neural Networks approach in [11] and the same obtained factor using Spatial Reasoning and Temporal Stack Learning [12] is better than the one gotten in [3]. Furthermore, the use of Deep Progressive Reinforcement Learning for Skeleton-based Action Recognition [13] highlighted the advantages against VA-LSTM [3] while testing on UT-Kinect Dataset.

References

[1] R. Poppe, "Poppe, r.: A survey on vision-based human action recognition. image and vision computing 28(6), 976-990," *Image Vision Comput.*, vol. 28, pp. 976–990, 06 2010.

[2] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE MultiMedia*, vol. 19, pp. 4–12, April 2012. [Online]. Available: https://www.microsoft.com/en-us/research/publication/microsoft-kinect-sensor-and-its-effect/

[3] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," *CoRR*, vol. abs/1703.08274, 2017. [Online]. Available: http://arxiv.org/abs/1703.08274

[4] Yong Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1110–1118.

[5] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3d human action recognition," *CoRR*, vol. abs/1607.07043, 2016. [Online]. Available: http://arxiv.org/abs/1607.07043

[6] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1010–1019.

[7] 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 *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2012, pp. 28–35.

[8] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for rgb-d activity recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5344–5352.

[9] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," *arXiv preprint arXiv:1611.06067*, 2016.

[10] J.-F. Hu, W.-S. Zheng, L. Ma, G. Wang, and J. Lai, "Real-time rgb-d activity prediction by soft regression," in *European Conference on Computer Vision*. Springer, 2016, pp. 280–296.

[11] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7912–7921.

[12] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, "Skeleton-based action recognition with spatial reasoning and temporal stack learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 103–118.

[13] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive reinforcement learning for skeleton-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5323–5332.