Spatio-temporal network-based HAR analysis of RGB and depth data

Ch.Raghava Prasad

Department of Electronics and Communication Engineering, Koneru Lakshmaiah Education Foundation (KLEF), Deemed to be University, Vaddeswaram, Green fields, Guntur, Andhra Pradesh, India chrp@kluniversity.in,

DOI : 10.48047/IJFANS/V11/ISS8/351

Abstract

The problem of recognizing human actions has shifted from one of video processing to that of multi-model machine learning. This paper's goal is to use recurrence models to address the a forementioned issue. Here, a spatio-temporal model is built out of solely RGB video data by combining recurrent neural networks (RNN) with spatial convolutional neural networks (CNN). The model employs a variant of recurrent neural networks known as long short-term memory (LSTM) networks. To encapsulate the time series of the convolutional features extracted by CNNs, LSTM networks are used. In a series of videos, LSTMs study the evolution of action features over time. Results have demonstrated that temporal traits, when combined with spatial ones, are the most important.

1.Introduction

One of the most challenging jobs is recognizing the human body through a machine, as it is the most diverse biomechanical structure in the history of human evolution. Early human action detection techniques rely heavily on data from videos with blank or minimally complex backgrounds. However, in the recent decade, thanks to developments in camera technology, there has been a tremendous rise in the availability of multi modal datasets, making it the most widely examined research subject. In most cases, the large amount of data needed to properly train a CNN based on RGB video results in a high computational cost. When compared to more conventional machine learning approaches, deep learning models provide superior representational power and the most accurate projections of potential performance. To collect the multi-modal data, researchers are turning to Kinects and Intel real sense cameras. In addition to RGB color, Depth, and skeletal datasets, Kinect is the most popular sensor for generating data of this type. The models that take into account depth are described below. Distances between objects in a video or image captured by the Kinect sensor are represented visually as pixels. In [1], a deep learning model for HAR was used to generate depth motion maps (DMM). However, DMMs did not adequately capture the temporal dynamics. To get over this drawback, the depth videos are represented in three distinct kinds of dynamic history images [2]. Inputs to three convolutional neural networks (CNNs) with score fusion included depth maps of the complete body, individual body sections, and joints. It was demonstrated that three-stream CNN architectures with score fusion could benefit from employing dynamic depth, dynamic depth normal, and dynamic depth motion normal representations [3]. In [4], however, LSTMs effectively alleviate the memory bottleneck that plagued CNN-based depth algorithms. The proposed approach utilizes the spatial and temporal features of the RGB and depth modalities to acquire knowledge. The effort is motivated by the LSTM's proven ability to learn from skeletal joint data.

2. Modeling an Active Network with a CNN and an LSTM

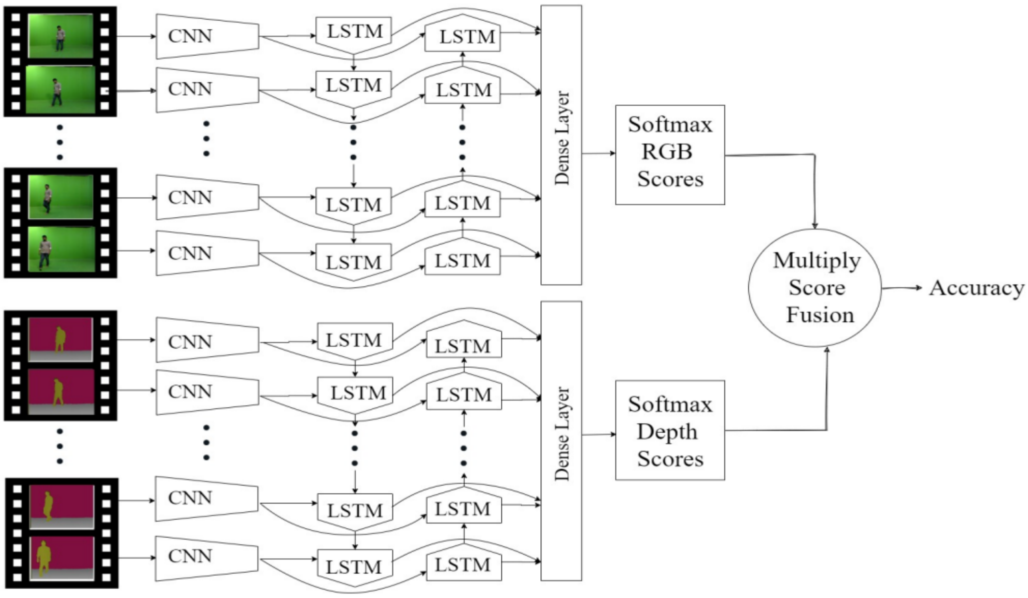

Fig. 1: The Proposed Framework: The stacked CNN – LSTM network

Skeletal data has limitations, however these can be mitigated by using data preparation methods. While data processing techniques are discussed in other chapters, they are not applied to any data in this one. In this section, we provide a CNN-LSTM architecture that can classify human behaviors from several streams of color and depth information. Fig. 1 depicts the overall CNN-LSTMs model architecture that underpins the proposed model. Figure 2 depicts the framework for the extraction of spatial features from RGB video frames. The CRGBe is a multi-layered ensemble of streams, consisting of 16. Six convolutional plus ReLu layers, three maximum pooling layers, and a flatten layer make up the ten-layer depth used across all streams.Seven-by-seven-by-five and five-by-three-by-three filter kernels were chosen for this study.

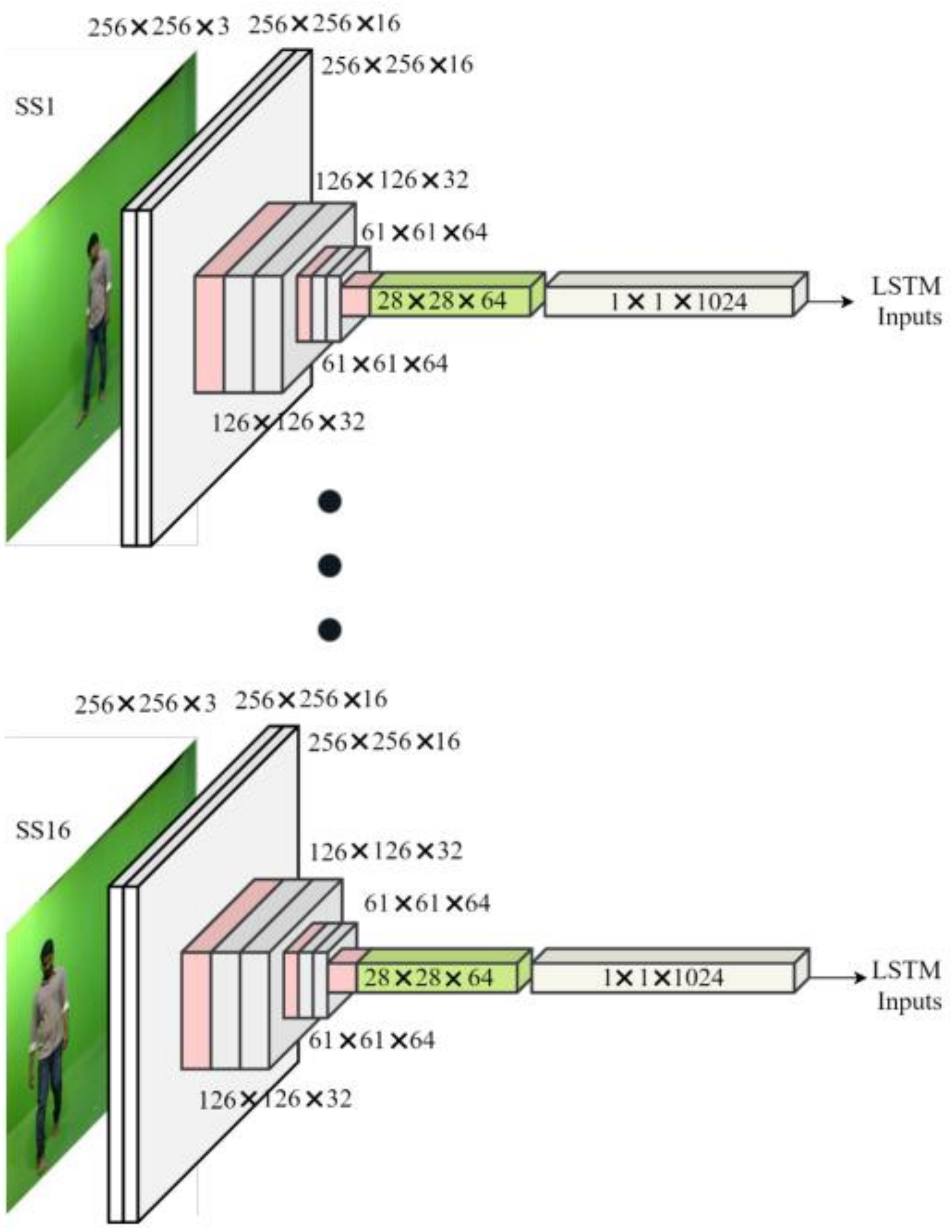

Fig. 2: The CRGBe ensemble for extracting spatial features from RGB action video frames

When training for a sequence labeling problem, like video-based action identification, it is helpful to have access to both historical and prospective inputs at the same time. As can be seen in Figure 3, this has already been accomplished with the use of bidirectional LSTM networks.

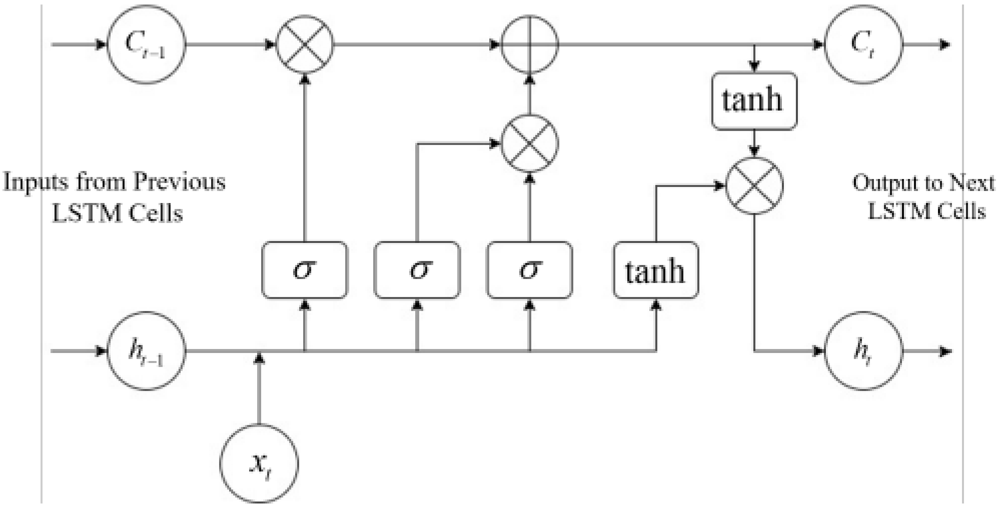

Fig. 3: A Single LSTM Cell architecture

3.Conclusion

In this study, we present an improved model called CNN-LSTM, which combines the advantages of both spatial convolutional networks and temporal recurrence models. A convolutional neural network (CNN) - long short-term memory (LSTM) model is used to train and test the suggested odel. It has also been discovered that this model's training procedure has a delay, which is a significant drawback. This lag is caused by the many trainable parameters in the chained Bi - directional LSTM units. In order to provide output representations of time series with no breaks, this sequence of LSTM blocks is necessary.

References

1.M. Al-Faris, J. Chiverton, Y. Yang, and D. Ndzi, “Deep learning of fuzzy weighted multi-resolution depth motion maps with spatial feature fusion for action recogni tion,” Journal of Imaging, vol. 5, no. 10, p. 82, 2019.

2. A. Snoun, N. Jlidi, T. Bouchrika, O. Jemai, and M. Zaied, “Towards a deep human activity recognition approach based on video to image transformation with skeleton data,” Multimedia Tools and Applications, vol. 80, no. 19, pp. 29 675–29 698, 2021.

3. Z. Tu, W. Xie, Q. Qin, R. Poppe, R. C. Veltkamp, B. Li, and J. Yuan, “Multi stream cnn: Learning representations based on human-related regions for action recognition,” Pattern Recognition, vol. 79, pp. 32–43, 2018.

4. S. Arif, J. Wang, T. Ul Hassan, and Z. Fei, “3d-cnn-based fused feature maps with lstm applied to action recognition,” Future Internet, vol. 11, no. 2, p. 42, 2019.

5. I. Afolabi, T. Taleb, K. Samdanis, A. Ksentini, and H. Flinck, "Network slicing and softwarization: A survey on principles, enabling technologies, and solutions," IEEE Communications Surveys Tutorials, vol. 20, no. 3 pp. 2429–2453, March 2018.

6. A. Kaloxylos, "A survey and an analysis of network slicing in 5G networks," IEEE Communications Standards Magazine, vol. 2, no. 1, pp. 60–65, April 2018.

7. S. Zhang, W. Quan, J. Li, W. Shi, P. Yang, and X. Shen, "Air-ground integrated vehicular network slicing with content pushing and caching," IEEE Journal on Selected Areas in Communications, vol. 36, no. 9, pp. 2114–2127, August 2018.

8. C. Campolo, A. Molinaro, A. Iera, and F. Menichella, "5G network slicing for vehicle-to-everything services," IEEE Wireless Communications, vol. 24, no. 6, pp. 38–45, December 2017.

9. Afify AA. Similarity solution in MHD: effects of thermal diffusion and diffusion thermo effects on free convective heat and mass transfer over a stretching surface considering suction or injection. Commun. Nonlinear Sci. Numer. Simul., 14:2202–2214, 2009.

10. Turkyilmazoglu M. Multiple solutions of heat and mass transfer of MHD slip flow for the viscoelastic fluid over a stretching sheet. Int. J. Therm. Sci., 50:2264–2276, 2011.

11. Turkyilmazoglu M. Dual and triple solutions for MHD slip flow of non- Newtonian fluid over a shrinking surface, Comput.Fluids,70:53–58, 2012.

12. Rashidi MM, Erfani E. Analytical method for solving steady MHD convective and slip flow due to a rotating disk with viscous dissipation and Ohmic heating. Eng. Comput., 29:562–579, 2012.