2022 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE) | 979-8-3503-1156-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/WIECON-ECE57977.2022.10150907

Role Identification from Human Activity Videos using Recurrent Neural Networks

Anam Arshad
Department of Computer Science & Engineering
International Institute of Information Technology (IIIT)
Naya Raipur, India
anam21300@iiitnr.edu.in

Vivek Tiwari
Department of Computer Science & Engineering
International Institute of Information Technology (IIIT)
Naya Raipur, India
vivek@iiitnr.edu.in

Mayank Lovanshi
Department of Computer Science & Engineering
International Institute of Information Technology (IIIT)
Naya Raipur, India
mayank@iiitnr.edu.in

Rahul Shrivastava
Department of Computer Science & Engineering
Sagar Institute of Science, Technology & Research
Bhopal, India
rahul.vidishaa@gmail.com

***Abstract*—Recognizing the underlying roles in mutual activity is more informative. Role identification has the potential to improve wide range of applications, of activity recognition from safety and security to healthcare. In recent research, for role identification, work is done to identify roles by capturing the knowledge of body parts from an image. This work is complex and not sufficient to take input as English sentences and capture the sequencing and relationship between words. There is a need for simple work which could use recent technologies like Recurrent Neural Networks to capture the recurrent nature of sentences to identify roles. The contribution of this work is proposing a Computational Long Term Memory model where sentences are stored as features and given to a Recurrent Neural Network to identify roles. The appropriate dataset is not available for role identification using sentences. In this view, this work attempt to develop a new custom dataset. The proposed model is tested on accuracy using various Recurrent Neural Networks like Recurrent Neural Network (RNN), Long Short Term Memory (LSTM), Gated Recurrent Units (GRU) etc. The LSTM model gave effective accuracy of 60% on the small custom dataset.**

***Index Terms*—Role identification, Recurrent Neural Networks, Long Short Term Memory, Gated Recurrent Units.**

I. Introduction

Human Activity Recognition can have an impact on a wide range of applications including surveillance, forecasting, offloading sensor data, etc [1]. Recognizing atomic activities such as walking, standing, cycling, etc is a very popular exemplar for activity recognition research [1]. Examples of atomic activities are shown in Figure 1. Most of this research focuses on the discrete behavior of a single individual rather than what other individuals are doing in the same environment [2].However, it is imperative to scrutinize the collective behavior of how every individual interacts with each other in the same environment.

Many activities are associated with the behavior of each individual in the environment such as giving, taking, pushing, punching, etc. These activities are called reciprocal activities. They are shown in Figure 2. Recognition of these activities is

Fig. 1. Examples of Atomic activities [3]

an arduous task. For instance, if a 'Boy' is giving something to a 'Girl', then in mutual activity recognition, it becomes difficult to identify whether "taking" or "giving" is the activity, as both these activities are correlative to each other [4]. Consequently, Role identification is important [4].

Role identification of each individual in the same environment is more elucidate. Role identification plays a very significant role in recognizing reciprocal activities [4]. In the example, a 'Boy' giving something to a 'Girl', to recognize the activity to be "giving" or "taking", we need to identify the roles of each entity. Recognizing the entities and their respective roles is very important to capture the collective behavior of all the individuals in the same environment [4]. Role identification is only viable within the context of a specific reciprocal activity [4].

Recognizing the abstract actions involving person-person contact is the main focus in view of current improvements [4]. Providing training for role identification in the context of reciprocal engagement is challenging since it necessitates role labelling for each person [4]. In this article, a Computational

979-8-3503-1156-3/22/31.00 ©2022 IEEE

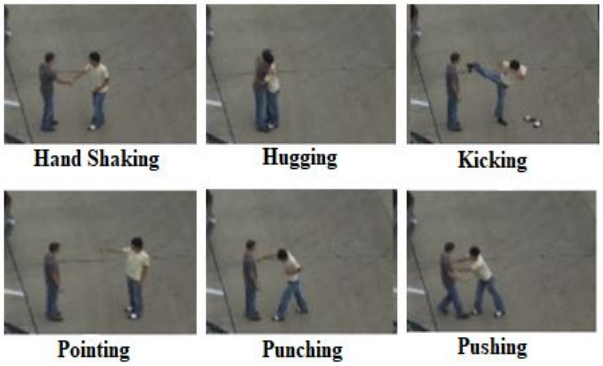

Fig. 2. Examples of Reciprocal activities [5]

Long Term Memory model is created where sentences are stored in form of features and given to different types of Recurrent Neural Networks to identify roles. The contribution of the work can be summarized as follows:

- A novel Computational Long Term memory model is proposed to identify the underlying roles.
- A novel dataset is created specifically for performing role identification.
- Effective way to generate the input feature vector is proposed.
- To demonstrate the significance of our work, Recurrent Neural Networks like RNN, LSTM, and GRU have been applied to newly created dataset to capture the relationship existing between sequence of words.

In this work, Section II describes the literature survey of mutual activity recognition and role identification. Section III describes the novel proposed architecture in more detail. Section IV describes the evaluation of our proposed model using various evaluation metrics. Section V describes the conclusion and future scope of our work.

II. Related Work

Mutual activity recognition is an active area of research and works in this field is profuse, but still, it is a big challenge to recognize those activities where multiple bodies interact [3]. There are many existing algorithms and solutions for mutual activity recognition. These solutions consist of approaches like capturing the latent Spatiotemporal dependencies among body regions and persons [6]. These approaches are effective and give superior performance as compared to other numerical stable solutions. However, these solutions cannot capture the underlying roles.

The development of extremely advanced capture tools that can determine the depth and 3D coordinates of the human body [7], they have been widely utilised for pose estimation [8]. The pose estimation methods can be used directly to identify the collective behavior of all the individuals in mutual activity recognition [9]. Additionally, there have been a number of developments in mutual activity recognition, including a framework for graphical inference that can be used to learn the human species' consistent behaviour [10]. Furthermore, with the known benefit of combining Recurrent Neural Networks (RNN) with Convolutional Neural Networks (CNN) [11] to capture the temporal dynamics of poses, since then they became the backbone for mutual activity recognition [12]. In the realm of pose-based action recognition, well-known temporal models like RNN/LSTM and spatial model convolutional models (CNN/GCN) have long dominated the field. [13]. Some solutions combined the use of LSTM and confidence energy to compute p-values of the solutions for estimating minimum confident energy [14]. The solutions given are reliable but are too complex to be used for mutual activity recognition and role identification.

Advancement in the field of group activity recognition [15] introduced many new methods. These methods include a graphical model combined with neural networks for group activity recognition [16]. Some approaches employ the dynamic programming concept in conjunction with an AND-OR graph to identify events and assign roles to the participants in those events [17]. However, these solutions are not sufficient for recognizing group activity. Although there is a majority of prior work with propitious results using RNNs and CNNs for mutual activity recognition, they still lack in detecting the roles in activity recognition.

Some methods used the knowledge of the human body parts to coordinate role detection by learning the hidden Spatiotemporal dynamics among features [4]. These solutions can identify roles in Spatiodynamics features where input is given in form of an image [4].The majority of the solutions in the area of mutual activity recognition and role identification are primarily controlled by models that encode images and videos as numeric features, omitting to recognise mutual reciprocal activities.Thus, to replace the existing popular architectures, an effective paradigm for reciprocal activity recognition must be created.

There are a wide variety of datasets available for mutual activity recognition. NTU [18] and SBU [8] are the most prominent datasets available for mutual activity recognition. In addition to the COLLECTIVE ACTIVITY dataset [14] and VOLLEYBALL, dataset [6] are also available for group activity recognition. The majority of prior work for mutual activity recognition has been done using these datasets but they are not sufficient for identifying roles. MARD dataset [4] is specifically designed for role detection using Spatiotemporal dynamics among human body parts. Although, MARD dataset is specifically for role detection, it cannot identify the roles from textual data. Thus there is a need of creating a new dataset that can capture the sequencing between words to identify roles.

III. Proposed Work

This section proposes a Computational Long Term Memory model where roles are identified from English sentences. The proposed system is a Long Term Memory model architecture where sentences are encoded as features and given to various Recurrent Neural Networks to identify roles in form of an

output vector having the target features. The workflow of the proposed architecture is given in Figure 3. The proposed architecture goes through three stages. In the first stage it takes input in form of sentences. Sentences are tokenized and converted into words. Here, a customised POS tagging is employed in this piece of work to replace only the chosen term with its broad category or class.Partial Parts of Speech (POS) tagging is applied to the tokenized words to preserve all those words which have more influence in role identification.For example terms like ”what,” ”who,” ”why,” and ”whom,” for instance, can be substituted by the POS tagger's ”WH” category because each one is in charge of identifying various roles. As a result, these words are utilised just as they are now.The frequency of occurrences of each distinct word is then determined from all of the sentences. The total number of words in each sentence is then determined and a corpus of sentences is created from these pre-processed words.

In the second stage from the corpus of sentences unique words are extracted.These unique words are encoded as vectors using One Hot Encoding.These Encoded vectors are reshaped into the size of number of sentences,maximum size of each sentence,number of unique words.This reshaped vector is then passed through the third stage.

In the third stage various types of Recurrent Neural Networks like Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) are used to capture the relationship that exists between sequence of words.The output given by the training model is also in form of vectors representing the target features. These target features are the identified roles.

The detailed description on how each model is working in third stage is given below:

A. Recurrent Neural Network(RNN)

A simple Recurrent Neural Network is used to extract the output vector for each word in our proposed model.Recurrent Neural Networks are intelligent enough to take input in form of sentences and extract features from them implicitly [19]. RNN takes input in form of vectors which represent the words encoded as features and gave the output in form of vectors representing the target columns.

B. Long Short Term Memory(LSTM)

A long Short term Memory Neural Network is also used to extract the output vector for each word in our proposed model.In terms of capturing the recurrent nature of sentences LSTM perform the best [20].They are a Recurrent Neural Network extension with a different hidden layer [20]. The LSTM's hidden layers are gated cells with four layers that produce the cell's output as well as its state [21]. LSTM also takes input in form of vectors which represent the words encoded as features and gives output in form of vectors representing the target features.

C. Gated Recurrent Units(GRU)

The Gated Recurrent Units is also used to extract the output vector for each word in our proposed model. They are the extensions of LSTM with the only difference of having two gates while LSTM has three gates [22].GRU is less complex as compared to LSTM. GRU is preferred over LSTM for small datasets [23]. GRU also takes input in form of vectors and gives output in form of vectors representing the target features.

IV. Experimental Results

This approach is evaluated by the newly created custom Dataset for role identification. The samples from the datset is shown in Table I. The new dataset was designed especially for identifying roles. Below are detailed explanations of the datasets.

A. Dataset

The mutual role identification dataset is the custom dataset, specifically for the envisaged task of mutual role identification. The dataset takes English sentences along with 16 target features. The dataset is small in size with a length of 58 sentences. The dataset is created extensively for capturing the recurrent nature of sentences. Although, MARD dataset [4] is there for role identification, it is not sufficient to capture the sequencing of words in the sentences. MARD dataset can be used for role identification by capturing the Spatiotemporal dynamics of human body parts where input is given in form of an image. However, to perform role identification in the proposed work, capturing the sequence of words is necessary. Thus, this custom Dataset is suitable for the proposed work. The description of the target columns of the the custom Dataset is shown in Table II.

B. Comparison Between Recurrent Neural Networks

As discussed earlier various types of Recurrent Neural Networks like Recurrent Neural Networks (RNN), Long Short term memory(LSTM), and Gated Recurrent Units(GRU) were used to capture the recurrent nature of sentences present in the custom Dataset. These Neural Networks were used to encode each word present in the dataset as a feature so that they represent sequencing. The output vector extracted for words is a target vector representing each role present in the 16 target columns.

In evaluating the experimental results, we use multiclass accuracy (the proportion of accurate predictions) as an evaluation parameter. Furthermore, the experiment conducted for role identification on the newly created custom Dataset is processed by taking the same number of samples as present in the dataset. The dataset has 58 example sentences; 47 of them are used for training and the rest for testing. The results of our experiment are tabulated in Table III. The ratio of the number of accurate predictions to the total testing samples serves as the evaluation indicator for role identification accuracy. A detailed comparison of various parameters is given in the subsections below:

1) Accuracy: The accuracy is calculated for word sequencing by each Neural Network, which is reported highest by LSTM. LSTM gave the highest training accuracy of 80% and the highest testing accuracy of 60%, while simple RNN gave

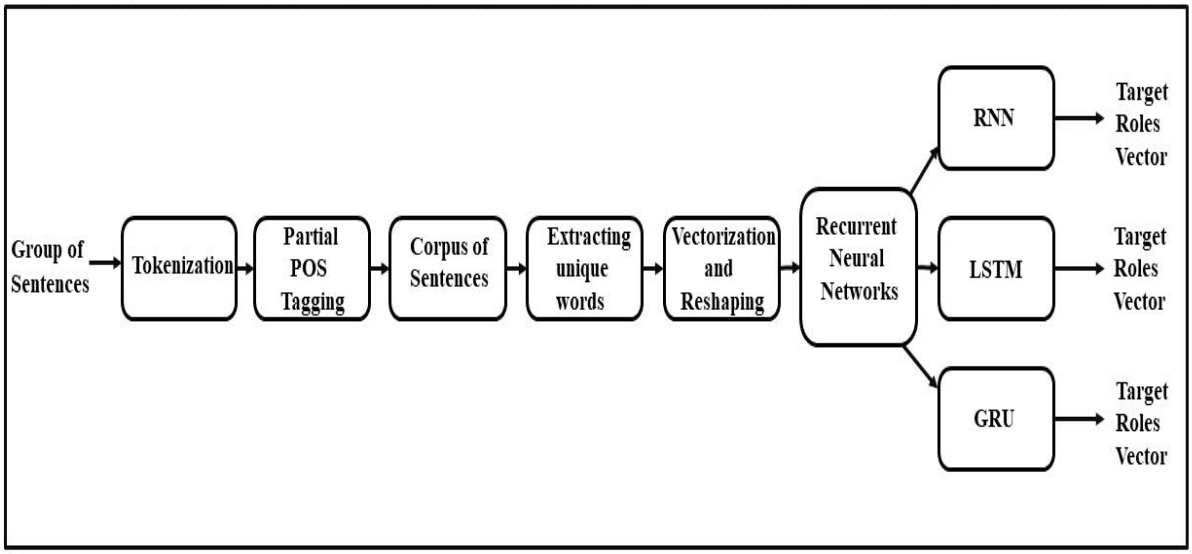

Fig. 3. Workflow of the Proposed Architecture

TABLE I
SAMPLES FROM THE NEWLY CREATED SAMPLE DATASET

Sentences	what has been verb	who has verb	who has been verb	time	what verb
”What has ram given to sita”	1	1	1	0	0
”Who has given book to sita”	1	1	1	0	0
”What has been given to sita”	1	1	1	0	0
”Did you call me yesterday”	0	1	1	1	0
”From whom do you take books”	0	0	0	0	1
”To whom has Ram given books”	1	1	1	0	0
”Ram has hitted Shyam”	0	1	1	0	0
”Ram has hitted shyam by stone”	1	1	1	0	0
”Ram has given all books”	1	1	0	0	0
”Sita has taken books from library”	1	1	0	0	0

a training accuracy of 75% and a testing accuracy of 40%. In addition, GRU gave a training accuracy of 50% and a testing accuracy of 41%. The results of training accuracy and testing accuracy for each model are shown in Table II.

2) Loss: For instance, on observing both training and testing accuracy we find that in some models there is a difference in training and testing accuracy. This indicates that our model is overfitting. The main reason for overfitting is the small size of the custom Dataset. Thus, there is a need to increase the size of our dataset. Hyperparameter tuning may also help us in improving the performance of the model. The graphs of training loss and testing loss for all the models are shown in Figure 4, Figure 5 and Figure 6. From the graph, we observe that Long Short term memory(LSTM) is outperforming as compared to other models. Thus in the future by increasing the dataset, LSTM will perform better both in terms of increased accuracy and decreased loss, as compared to RNN and GRU.

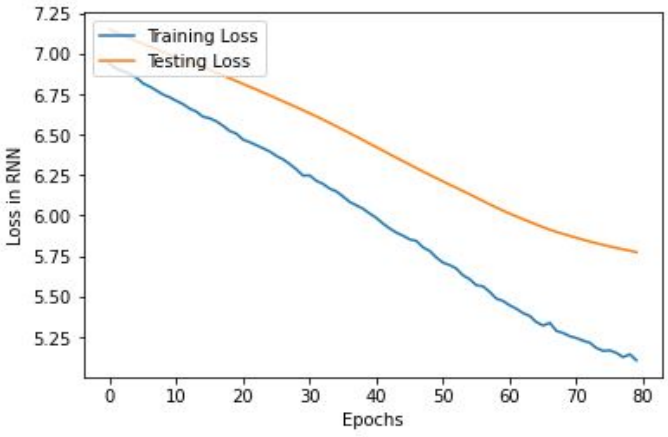

Fig. 4. Training loss vs Testing for RNN

V. CONCLUSION AND FUTURE SCOPE

In this work, a Computational Long Term Memory model is created to identify roles. Primarily, the model stores sentences from which words are extracted and given as input features. Based on the model, these input features are passed through various Neural Networks to extract sequencing between words.

TABLE II
DESCRIPTION OF TARGET COLUMNS IN THE DATASET

Target Columns	Description
WHAT_HAS_VERB3	If sentence has what +Third form of the verb.
WH0_HAS_VERB3	If sentence has who +Third form of the verb.
WHOM_HAS_VERB3	If sentence has whom +Third form of the verb.
WHAT_VERB	If sentence has what +First form of the verb.
WHO_VERB	If sentence has what + First form of the verb.
WHOM_VERB	If sentence has what +First form of the verb.
HOW_VERB	If sentence has what +First form of the verb.
Time	If Time is described in the sentence.
REASON_OF_VERB	If sentence gives a reason for the verb.
Event	If Event is described in the sentence.
PLACE_OF_EVENT	If Place of the event is described in the sentence.
FROM_WHERE_HAS_BEEN_VERB3	If Sentence has from where has been + Third form of the verb.
WHERE_HAS_BEEN_VERB3	If Sentence has where has been + Third form of the verb.

TABLE III
PERFORMANCE COMPARISON WITH TRAINING ACCURACY ON CUSTOM DATASET

Recurrent Neural Networks	Training Accuracy	Testing Accuracy
RNN ([19])	75.00%	40.00%
LSTM ([21])	80.04%	60.04%
GRU ([23])	50.00%	41.88%

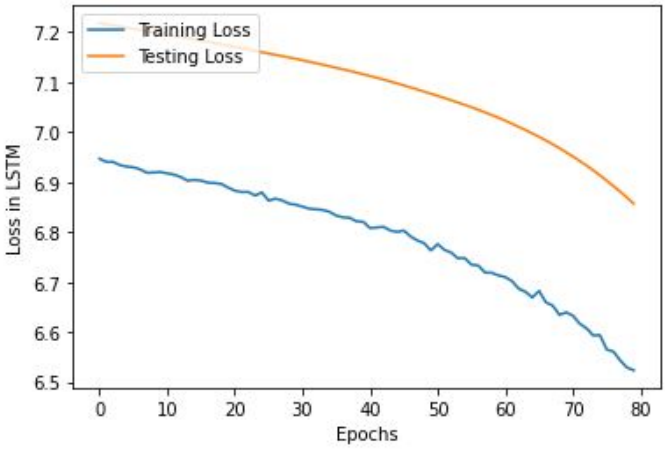

Fig. 5. Training loss vs Testing for LSTM

The efficiency of the model was tested using role identification experiments on the newly produced dataset known as the custom Dataset utilising recurrent neural networks. The recurrent nature of sentences was captured using recurrent neural networks such as recurrent neural networks (RNN), long short term memory (LSTM), and gated recurrent units. LSTM performs better as compared to others with an accuracy of 60% testing accuracy, while RNN gave testing accuracy of 40% and GRU gave testing accuracy of 41%. From the graphs of training loss and testing loss of each model shown in Figure 4, Figure 5 and Figure 6 and Training accuracy and Testing accuracy shown in Table II, we conclude that performance of LSTM is the best as compared to RNN and GRU because the technique employed to create feature vectors actually creates

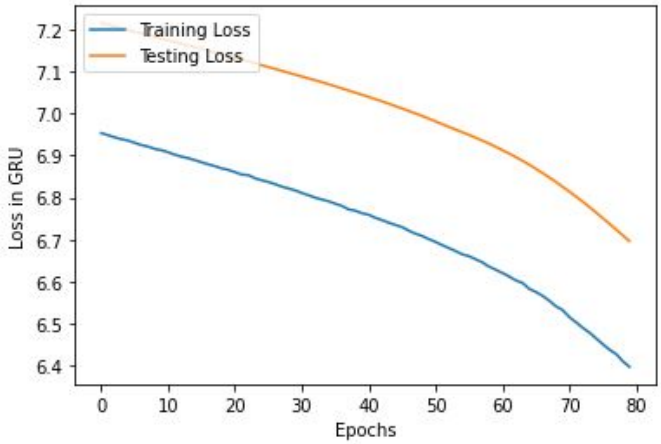

Fig. 6. Training loss vs Testing for GRU

feature vectors that reflect lengthy word sequences.

The proposed Long Term Memory model can be further improved by experimenting with the fusion of various RNN models. The model can also be improved by performing hyper-parameter tuning. In addition, the accuracy of the model can also be improved by increasing the size of the custom Dataset. Moreover, another Neural network can also be created to bind the role and entity vectors together to extract the mutual activity from it. Future research in mutual role identification has many potential avenues yet to be explored.

REFERENCES

[1] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3595–3603, 2019.

[2] Wongun Choi and Silvio Savarese. Understanding collective activitiesof people from videos. *IEEE transactions on pattern analysis and machine intelligence*, 36(6):1242–1257, 2013.

[3] Michalis Vrigkas, Christophoros Nikou, and Ioannis A Kakadiaris. A review of human activity recognition methods. *Frontiers in Robotics and AI*, 2:28, 2015.

[4] Rahul Shrivastava, Vivek Tiwari, Swati Jain, Basant Tiwari, Alok Kumar Singh Kushwaha, and Vibhav Prakash Singh. A role-entity based human activity recognition using inter-body features and temporal sequence memory. *IET Image Processing*, 2022.

[5] Chaolong Zhang, Yuanping Xu, Zhijie Xu, Jian Huang, and Jun Lu. Hybrid handcrafted and learned feature framework for human action recognition. *Applied Intelligence*, pages 1–17, 2022.

[6] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Higcin: Hierarchical graph-based cross inference network for group activity recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[7] Arash Vahdat, Bo Gao, Mani Ranjbar, and Greg Mori. A discriminative key pose sequence model for recognizing human interactions. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1729–1736. IEEE, 2011.

[8] Kiwon Yun, Jean Honorio, Debaleena Chattopadhyay, Tamara L Berg, and Dimitris 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*, pages 28–35. IEEE, 2012.

[9] Xiangbo Shu, Liyan Zhang, Guo-Jun Qi, Wei Liu, and Jinhui Tang. Spatiotemporal co-attention recurrent neural networks for human-skeleton motion prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3300–3315, 2021.

[10] Jinhui Tang, Xiangbo Shu, Rui Yan, and Liyan Zhang. Coherence constrained graph lstm for group activity recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[11] Meenakshi Choudhary, Vivek Tiwari, and Swati Jain. Person re-identification using deep siamese network with multi-layer similarity constraints. *Multimedia Tools and Applications*, 81(29):42099–42115, 2022.

[12] Xin Li and Mooi Choo Chuah. Sbgar: Semantics based group activity recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2876–2885, 2017.

[13] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1971–1980, 2016.

[14] Tianmin Shu, Sinisa Todorovic, and Song-Chun Zhu. Cern: confidence-energy recurrent network for group activity recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5523–5531, 2017.

[15] Ashish Singh Patel, Ranjana Vyas, OP Vyas, Muneendra Ojha, and Vivek Tiwari. Motion-compensated online object tracking for activity detection and crowd behavior analysis. *The Visual Computer*, pages 1–21, 2022.

[16] Rui Yan, Xiangbo Shu, Chengcheng Yuan, Qi Tian, and Jinhui Tang. Position-aware participation-contributed temporal dynamic model for group activity recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[17] Tianmin Shu, Dan Xie, Brandon Rothrock, Sinisa Todorovic, and Song Chun Zhu. Joint inference of groups, events and human roles in aerial videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4576–4584, 2015.

[18] Xiangbo Shu, Liyan Zhang, Yunlian Sun, and Jinhui Tang. Host–parasite: Graph lstm-in-lstm for group activity recognition. *IEEE transactions on neural networks and learning systems*, 32(2):663–674, 2020.

[19] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.

[20] Davi Guimarães da Silva, Marla Teresinha Barbosa Geller, Mauro Sérgio dos Santos Moura, and Anderson Alvarenga de Moura Meneses. Performance evaluation of lstm neural networks for consumption prediction. *e-Prime-Advances in Electrical Engineering, Electronics and Energy*, 2:100030, 2022.

[21] Md Sanzidul Islam, Sadia Sultana Sharmin Mousumi, Sheikh Abujar, and Syed Akhter Hossain. Sequence-to-sequence bangla sentence generation with lstm recurrent neural networks. *Procedia Computer Science*, 152:51–58, 2019.

[22] Rahul Dey and Fathi M Salem. Gate-variants of gated recurrent unit (gru) neural networks. In *2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, pages 1597–1600. IEEE, 2017.

[23] Shudong Yang, Xueying Yu, and Ying Zhou. Lstm and gru neural network performance comparison study: Taking yelp review dataset as an example. In *2020 International workshop on electronic communication and artificial intelligence (IWECAI)*, pages 98–101. IEEE, 2020.