Human Action Recognition using CNN and LSTM-RNN with Attention Model

Kuppusamy. P, Harika. C

Abstract*: The recent advancements in artificial intelligence make the world into recognizing the objects, learning the environment, and predicting the forthcoming sequences. Emerging of embedded technology leads to decrease the cost of surveillance systems. The surveillance systems are capturing the environment and stored in memory. Machine learning is utilized for processing the data to aware of the scenario. This paper is considered the idea of using the video for recognizing the human action and behavior. Tis paper is proposed the integration of convolutional neural network and long short-term memory recurrent neural network for processing the video. The convolution processes the given input that produces the informative spatial features. The extracted features directed into long short-term module to generate temporal features. The feature maps of long short-term memory component fed into proposed attention element. It captures the highly valuable informative features in the frame of video. The actions are recognized from the informative features using softmax module. This model is used to recognize the human actions from video. The experimental results proved that proposed model performed better with accuracy.

***Index Terms*: attention model, behavior, CNN, human action, LSTM.**

I. INTRODUCTION

The recognition of human activity is one of the primary issues in real-time environment. At present, researchers have been working on this issue since it has received sensible attention in the computer vision. Human activity analysis is used in different industry and security applications such as intelligent surveillance system, etc. An individual action has recognized and provided good performance in recognizing process [1-2]. But, it is not recognizing the activity in multiple people environment. Because, most of the real-time scenario consists multiple people interactions or movement.

Context based human action recognition has been proposed that improved the conventional target -centered action recognition. Lan et. al. has presented an action recognition method using encoding the human interactions between multiple people [3]. Choi et. al. has exploited spatial-temporal features to observe the adjacent human actions [4]. Most of an existing human action recognition methods exploits the people as context information without considering the background context information, location in which the activity is taken place, the locality of human within the scene, etc. [5].

Revised Manuscript Received on June 12, 2019

Kuppusamy. P, Department of Computer Science, Madanapalle Institute of Technology & Science, Madanapalle, India.

Harika. C, Department of Computer Science, Madanapalle Institute of Technology & Science, Madanapalle, India.

The existing approaches have exploited the context information either as input features to random forest classifier [4] and support vector machine [6] or fused the context information over probabilistic methods such as conditional random fields. Deep Neural Network (DNN) models have the processing influence in fusing multi-sources of context information. The multi-layer deep architecture is capable of handling the probabilistic reasoning, hidden neurons integration to fuse complex level representations of the raw input feature [7].

Smart cities are emerging in worldwide since the growing of population and advancement in ICT tools. Smart homes are also increasing to handle the electronics and electrical equipment's over the mobile phone. The primary methods of human action recognition are vision and sensor based monitoring. The sensor-based action recognition and prediction is widely used in intelligent environments [8, 9]. The vision-based action methods are led to make the user concern about the privacy [10].

II. RELATED WORK

The emerging of sensor network technology leads to exploit the sensor-based action recognition and monitoring in real-time applications. The sensors are generating the sequential and time series data that changes the state parameters. These parameters value is processed using fusion, probabilistic or statistical models for recognizing the human activity recognition. It consists two different techniques such as data-driven and knowledge-driven for recognizing the human action in sensor-based environment. First method data-driven is utilizing the data mining and machine learning approaches to observe the action [11].

In supervised machine learning, human actions are learned and recognizing the activity from gathered sensor data. In Data-driven method, labelled huge datasets are utilized to learn the human actions through training phase with various classifiers. The past domain knowledge is utilized in knowledge-driven methods that avoids the manual labeling process for training datasets. Ismail and Hassan have presented the approach to extract the action features from the human body that is used for recognizing many actions [12]. Chen et al. proposed activity recognition method using prior domain knowledge of experts [13]. An ontology-based methods have represented the features explicitly to facilitate reusability, interoperability and portability [14-16]. Human mobility patterns and device utilization is predicted using compression,
hidden-markov model, and sequence matching approach that enables the system to

recognize the human needs in surveillance environment. Human action is recognized in smart environments.

The two types of human behaviors models are intra-activity behavior defines how the human performs activity and inter-activity behavior defines the actions and activity sequences [17]. An action sequences are utilized for modeling the inter-activity human behavior that predicts the future behavior. It provides the fine-grained descriptors of human activity during extracting the features from particular sensor-based environment. Human behavior is the combination of various components that make a complex structure. Action is defined as simple and short sensible movement of organ in a body such as keeping a book on the table, close the door, clapping, etc. Human activity is defined as the sequence of complex conduct of movements such as playing a foot-ball, watching a television, riding the bi-cycle, etc. Human behavior defines how the human done the activities at various situations.

This paper presents the inter-activity human behavior model that showed how actions are performed. The characteristic of sensor-based environment algorithm is flexible to process the data of intelligent environments. It is mapped the raw sensor input into actions features. Hoey et al. have been proposed the action recognition model that process the video and mapping the actions to evaluate the task of patients' hand-washing with dementia disease [18]. Kruger et al. have been presented the model that process the inertial measurement and mapping the actions to describe the activities through state-space computational models. The actions are clustered into various classes [19].

The Convolution Neural Network(CNN) is one of the model to extract spatial features of images using the convolution layers. These layers consist the orientation-sensitive filters [20] and also captures the temporal features from the video dataset [21]. The traditional Recurrent Neural Network (RNN) approach combined with Long Short Term Memory (LSTM) to enable the network to preserve various time sequential data with long-term dependencies [22].

III. HUMAN ACTION RECOGNITION USING LSTM-RNN

LSTM is exploited with an attention model to improve the performance in translation [23] and image tagging [24]. LSTM is provided the good performance in activity prediction from video. However, it does not consider the spatial correlation. Shi et al. have extended the CNN model with adding the more convolutions in the network that captures the spatial and temporal features from the video frames to improve the performance [25].

This paper proposes a novel architecture for human action recognition using LSTM based RNN from video frames. The proposed method contains CNN, LSTM and attention model. The convolution layer captures the spatial information. Consequently, LSTM layer captures the temporal feature information. The attention model is combined with LSTM that captures the important feature information of video that avoid the unwanted noise from the frames. It leads to improves recognizing performance of the LSTM based network. The output of LSTM is a vector that notifies temporal feature information of video frames.

Fig 1. shows the LSTM-RNN framework with convolutional features and attention model. The CNN is filtered the spatial information from each frame of video and LSTN-RNN is explored the temporal information among the various frames in video. Here, attention model is combined with LSTM-RNN. The training phase of the network model is used the video labels for action recognition. The CNN is captured the different spatial information such as curves, shapes, location, invariance, rotation invariance, etc.). Attention model is used to emphasis the moving objects than entire image or static background that decreases the effect background effect. It led to increase performance of designed network model.

The CNN is used for extracting the spatial information from the RGB video frames using the combination of convolution and pooling process. The weights are initialized randomly and bias is included as input in every hidden layer. The feature maps are generated from fully connected layer of CNN. Each feature map consists F x F x C dimension. F x F represents the feature vector size, and C represents number of kernels. The FCNN feature maps (fm) (i.e output of CNN layer) are presented in matrices for the video length V.

$$F_{CNN} = [f_m^1, f_m^2, \dots\dots, f_m^t, \dots f_m^V] \quad (1)$$

$$F_{CNN} \in \mathbb{R}^{F \times F \times C \times V} \quad (2)$$

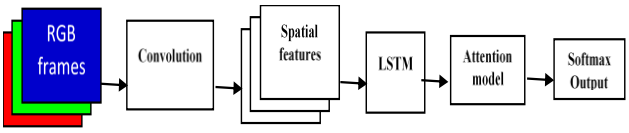

Fig 1. LSTM-RNN framework with attention model

The output of CNN is fed as input to the LSTM layer. The LSTM structure is showed in Fig. 2. LSTM consists the many components such as input gate, forget gate, input modulation gate and output gate. The input feature vector represents as xt, cell state as Ct, hidden state as ht and output state as Ot. The output is the tanh computation of hidden state. Hidden state values are computed using previous cell (Ct-1) and hidden (ht-1) states.

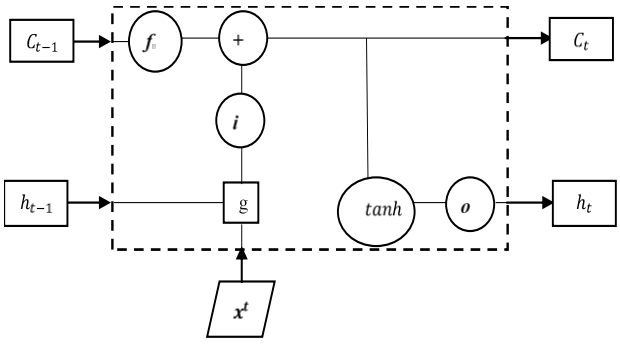

Fig 2. LSTM structure

$$o = sigmoid(w_o h_{t-1} + u_o x_t + b_o) \quad (3)$$
$$f = sigmoid(w_f h_{t-1} + u_f x_t + b_f) \quad (4)$$
$$i = sigmoid(w_i h_{t-1} + u_i x_t + b_i) \quad (5)$$
$$g = tanh(w_g h_{t-1} + u_g x_t + b_g) \quad (6)$$
$$h_t = o \odot tanh(C_t) \quad (7)$$
$$C_t = f \odot C_{t-1} + i \odot g \quad (8)$$

Cell state contains the input frame and memory information. It processes the frames and preserves the information of sequences over time. It is also capturing the long-term dependence information. LSTM consist the advantage that overcome the gradient vanish during the backpropagation over time. The following equations showed the LSTM computation model. h_{t-1} indicates the previous hidden state, w, u are input vector features and hidden state weights respectively. b_o, b_f, b_i, b_g denote the bias terms of output gate, forget gate, input gate and input modulation gate respectively. The symbol $\odot$ represents the Hadamard product. The long-term dependencies are captured with cell state and output.

The LSTM output is fed into temporal attention component that select the valuable information frames among all the frames in video. These valuable information frames are used to improve the action recognition performance. The attention model output is computed and generated the feature output.

The feature vector consists the informative data which is received from video frames. These informative details are observed from by the attention model. It is mentioned in equation (9). I_t denotes the intermediate output of the model that is computed using $tanh$ activation function. It is computed using the weight parameters w_{fc} of fully connected layer and LSTM output o_t, and bias b_{fc}.

$$I_t = tanh(w_{fc} o_t + b_{fc}) \quad (9)$$

Softmax is utilized for computing the classification of action from N number of important frames of the video. N represents the number of frames and W_t^V denotes the weight parameter to the tth component of softmax function. $Output_t$ represents the probability of the important information frame to decide the action recognition.

$$Output_t = \frac{exp(W_t^V u_t)}{\sum_{n=1}^{N} exp(W_t^V) u_t} \quad (10)$$

The backpropagation is proposed to tune the weights in the network to attain the performance in training phase. The cross entropy is proposed as a cost function J that is defined as follows,

$$J = -argmax \sum_{i=1}^{P} ground\ truth^i log(Output^i) \quad (11)$$

P stands for the number of action classes. $\boldsymbol{ground\ truth^i}$ denotes the true labelled action and $Output_t$ represents the predicted action label from the proposed network model.

IV. EXPERIMENTAL SETUP

The proposed network model is evaluated through experiments for recognizing the actions in video. The public dataset UCF Sports [26] is exploited for classifying the human actions. Each video file consists the various number of frames. This proposed method considered the group of 40 frames per video [27] that provide better performance. Tensorflow is used for experimenting the model and used the optimization model Adaptive Moment Estimation (Adam). The learning rate is fixed for weights tuning as 10-3. The data is normalized and batch size is given as 40.

V. RESULTS AND DISCUSSIONS

The proposed approach has been tested with UCF Sports dataset. The proposed method's experimental results have been presented that shows improvement of recognizing process using attention model with LSTM. It performed better using LSTM with attention model. The Table 1 denotes the results of various models and compared with proposed method. The proposed method is performed slightly better than the fully connected layer LSTM model with respect to using attention and without attention model component in the network

Table 1. Performance comparison

Approaches	Without attention model	With attention model
Fully connected LSTM	77.42%	79.00%
Proposed method	78.65%	82.03%

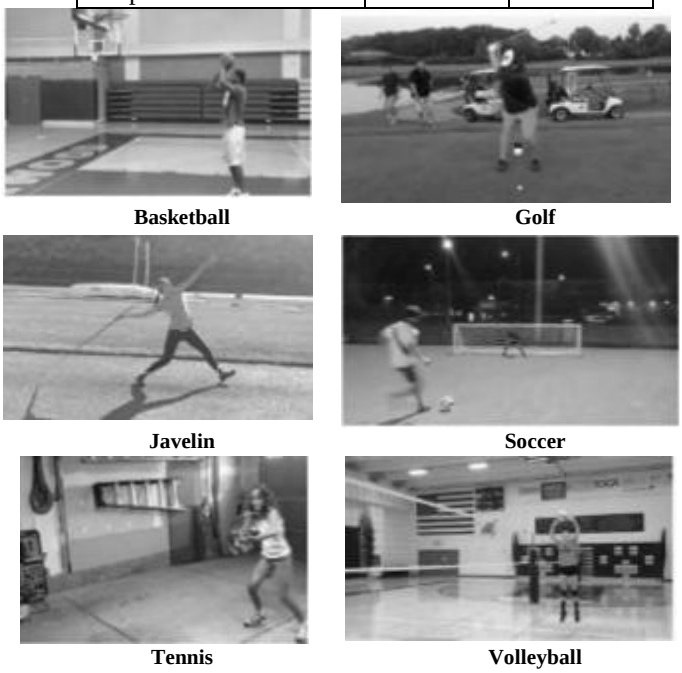

Fig 3. Action recognition using proposed model

Fig 3 shows the various action sequences. The proposed model is recognized the actions using the LSTM based RNN model.

Table 2. Performance comparison with accuracy

Approaches	Accuracy
Spatio-temporal features [39]	85.60%
Binary CNN-Flow [38]	94.80%
SGSH[43]	90.90%
Proposed method	90.05%

The Table 2 shows the accuracy of various methods and compare with proposed model. The proposed model is performed 0.8% to 4% better than two approaches. However, it is underperformed than the binary CNN-Flow method. This results proves the model performs slightly better than other methods.

VI. CONCLUSION

This paper is proposed the LSTM with attention model. The video is processed into sequence of frames. The RGB frames fed into CNN model to extract the informative features of frames such as curve, edge, color. The extracted features consist the shapes, outlines, coordinates, location orientation, etc. The LSTM is processed these spatial feature maps and results are produced as valuable temporal information features. The attention model component is proposed to identify the highly valuable informative features of video to recognize the action. The human actions are recognized using this proposed model. The results showed that proposed model performed slightly better than other methods. This work would be extended to produce good results to compare with binary-flow CNN model.

REFERENCES

1. D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition. Computer Vision and Image Understanding", 1152, 2011, pp. 224–241.
2. K. N. Tran, I. A. Kakadiaris, and S. K. Shah, "Part-based motion descriptor image for human action recognition. Pattern Recognition", 457, 2012, pp. 2562–2572.
3. T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," IEEE Transactions on Pattern Analysis and Machine Intelligence, 348, 2012, pp. 1549–1562.
4. W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition", Proc. IEEE International Conference on Computer Vision and Pattern Recognition, 2011.
5. X. Wang and Q. Ji, "Video event recognition with deep hierarchical context model", Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4418–4427.
6. K. N. Tran, A. Bedagkar-Gala, I. A. Kakadiaris, and S. K. Shah, "Social cues in group formation and local interactions for collective activity analysis", In VISAPP, 2013, pp. 539–548.
7. K. N. Tran, X. Yan, I.A. Kakadiaris, and S. K. Shah, "A group contextual model for activity recognition in crowded scenes". Proc. International Conference on Computer Vision Theory and Applications, 2015.
8. P. Kuppusamy, R. Kalpana and P. V. Venkateswara Rao, "Optimized traffic control and data processing using IoT", Cluster Computing, 2018, pp. 1-10. https://doi.org/10.1007/s10586-018-2172-5
9. P. Kuppusamy, P. Kamarajapandian, M. S. Sabari, and J. Nithya, "Design of Smart Traffic Signal System Using Internet of Things and Genetic Algorithm" In: Rajsingh E., Veerasamy J., Alavi A., Peter J. (eds) Advances in Big Data and Cloud Computing. Advances in Intelligent Systems and Computing, vol 645. Springer, 2018 pp. 395-403.
10. L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, Z. Yu, "Sensor-based activity recognition", IEEE Trans. Syst. ManCybern. Part C Appl. Rev, 42, 2012, pp. 790–808.
11. G. Ning, Z. Zhang, C. Huang, X. Ren, H. Wang, C. Cai and Z. He, "Spatially supervised recurrent convolutional neural networks for visual object tracking", In Proc. IEEE International Symposium on Circuits and Systems (ISCAS), 2017.
12. W. N. Ismail, and M.M. Hassan, "Mining Productive-Associated Periodic-Frequent Patterns in Body Sensor Data for Smart Home Care", Sensors (Basel), Vol. 17(5):952, 2017.
13. L. Chen, C. Nugent, M. Mulvenna, D. Finlay, X. Hong, and M. Poland, "A logical framework for behavior reasoning and assistance in a smart home", International Journal of Assistive Robotics and Mechatronics, vol. 9(4), 2008, pp. 20–34.
14. D. Riboni, and C. Bettini, "COSAR: Hybrid reasoning for context-aware activity recognition", Personal and Ubiquitous Computing, vol. 15, 2011, pp. 271–289.
15. L.Chen, C.D. Nugent, H.Wang, "A knowledge-driven approach to activity recognition in smart homes", IEEE Transactions on Knowledge and Data Engineering, vol.24, 2012, pp. 961–974.
16. H. Aloulou, M. Mokhtari, T. Tiberghien, J. Biswas, and P. Yap, "An adaptable and flexible framework for assistive living of cognitively impaired people", IEEE Journal of Biomedical and Health Informatics, vol. 18, 2014, pp. 353–360.
17. A. A. Chaaraoui, P. Climent-Pérez, and F. Flórez-Revuelta, "A review on vision techniques applied to human behavior analysis for ambient-assisted living", Expert Systems with Applications, vol.39, 2012, pp. 10873–10888.
18. J. Hoey, P. Poupart, A. V. Bertoldi, T. Craig, C. Boutilier, and A. Mihailidis, "Automated hand washing assistance for persons with dementia using video and a partially observable Markov decision process", Computer Vision and Image Understanding, Vol. 114, 2010, pp. 503–519.
19. F. Kruger, M. Nyolt, K. Yordanova, A. Hein, and T. Kirste, "Computational state space models for activity and intention recognition. A feasibility study", PLoS ONE, vol. 9(11), 2014: e109381, doi:10.1371/journal.pone.0109381.
20. K. He, X. Zhang, S. Ren and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification", IEEE International Conference on Computer Vision, 2015.doi/ 10.1109/ICCV.2015.123.
21. J. H. Yoo, "Large-scale Video Classification guided by Batch Normalized LSTM Translator", Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2017.
22. A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network", Machine Learning, 2018. arXiv:1808.03314
23. M.T. Luong, H. Pham, and C. D. Manning, ``Effective approaches to attention-based neural machine translation." [Online]. Available: https://arxiv.org/abs/1508.04025,2015.
24. S. Bai, and S. An, "A survey on automatic image caption generation", Neurocomputing, vol. 311, 2018, pp. 291-304.
25. X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong and W. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," Proc. Interantional Conference on Neural Information Processing Systems, 2015, pp. 802-810.
26. K. Soomro and A. R. Zamir, "Action recognition in realistic sports videos," In: Moeslund T., Thomas G., Hilton A. (eds) Computer Vision in Sports. Advances in Computer Vision and Pattern Recognition, 2015, pp. 181-208.
27. J. Y. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 4694-4702.

AUTHORS PROFILE

P. Kuppusamy received his Bachelor's degree in Computer Science and Engineering from Madras University, India, 2002, and Master's degree in Computer Science and Engineering from Anna University, Chennai, India, 2007. He has completed Ph.D in Information and Communication Engineering, Anna University, Chennai, 2014. At present, he is working as Professor in Madanapalle Institute of Technology & Science. He has published 30 research papers in leading international journals, 7 papers in IEEE international conferences and a book for Computer Science and Engineering. Currently, working on Machine learning, Internet of Things based research project with sensors, Raspberry Pi, Arduino for handling smart devices. He acts/acted also as a reviewer for various journals (e.g. Pervasive and Mobile Computing Elsevier, Journal of Computer Science from Science Publications) and International Conferences. He has organized research workshops, seminars based on Arduino controller with smart devices and FDPs. He is Member of IEEE, ISTE, IAENG and IACSIT. His specializations include Machine learning, Networks, Cryptography, Distributed computing and IoT. His current research interests are Machine Learning and Internet of Things.

C. Harika is M.Tech Scholar in Madanapalle Institute of Technology and Science, Madanapalle, Chittoor, Andhra Pradesh. She is currently doing research on Machine Learning.

.