Dynamic Gesture Recognition using LSTM and Tf-Pose for Human Action Analysis

Karthik Sundar C M[1*], Hitheash D[2] and SathyaDevi G[3]

[1, 2] UG Scholar, Department of Information Technology, St. Joseph's College of Engineering, OMR, Chennai – 119, Tamil Nadu, India.

[3] Assistant Professor, Department of Information Technology, St. Joseph's College of Engineering, OMR, Chennai – 119, Tamil Nadu, India.

karthiksundarcm@gmail.com

Abstract: Human Action Recognition (HAR) is a significant important application in many areas including surveillance systems, human-computer interfaces and even sports monitoring. In this research, we investigate pose estimation as the first step, followed by the use of Long Short-Term Memory (LSTM) networks for action Recognition in real-time. The system uses key points detection from human pose using the tf-pose estimation library and follows the LSTM model in order to capture temporal patterns of movement from the video frames. Such integration enables the model to encode complementary spatial and temporal information from video sequences. The textual description of the proposed framework is aimed at the action sequences like jumping, running, waving and is further compatible with real-time inference using live video streams. The proposed system is validated with several benchmark datasets where promising accuracy and efficiency are demonstrated against counterparts. Additionally, the model's capability of working on the skeletal data reduces the computational problem thus making it easy to deploy in low resource settings. Experimental outcome shows that the proposed approach yields superior performance to conventional action recognition methods, primarily due to its emphasis on the temporal behaviour of the human skeleton. Hence, our current study offers more meaningful information concerning increasing the efficiency and effectiveness of HAR systems using a deep learning framework.

Keywords: Human Activity Recognition, Body Pose Estimation, Recurrent Neural Network, TensorFlow Pose Estimation, Real-time Motion Recognition, Skeletal based Motion Recognition, Movement Temporal Characteristics, Video Processing, Animal, Gesture Recognition

1 Introduction

Human Action Recognition (HAR) is an important subfield of computer vision and machine learning and specifically dedicated to the identification and interpretation of human actions in video data. Its directional field is significant in intelligent monitoring, human-computer interface, and self-driving vehicles. The development of HAR systems has received importance because these are being used currently in operating systems. The rapid advancements seen in pose estimation, along with developments in the Recurrent Neural Networks (RNN), which includes the popular Long Short Term Memory LSTM networks, have further promoted the advancements of HAR through the observation of temporal change in motion sequence.

The HAR methods were essentially mainly fragmented in spatial and temporal characteristic features, but these could not handle with complex actions. In the case of modern trends, pose-based action recognition, body skeletal information is extracted using pose estimation technique such as tf-pose which estimates key points from frames

S. Kannadhasan et al. (eds.), Proceedings of the International Conference on Sustainability Innovation in Computing and Engineering (ICSICE 24), Advances in Computer Science Research 120,
https://doi.org/10.2991/978-94-6463-718-2_123

of the video. These key points are then passed through LSTM networks for real time action recognition that also incorporates both spatial as well as temporal aspects.

To address limitations, further advancements in the HAR systems have been firmed up by including the dynamics of attention-based mechanisms to project the focused areas at frame level, which in return has benefited in improving the recognition accuracy for fast or subtle movements. As the proposed approach, tf-pose is concatenated with LSTM networks for a real-time and low-complexity HAR application, which proves competitive performance despite the complex disturbances, such as noise or occlusion.

Lastly, combining the pose estimation and LSTM for HAR provides a valuable opportunity for developing related real-time activities. Therefore, future enhancements of the key point detection and the network configurations are expected to further the research.

2 Literature Survey

Human centred Action Recognition which has received a lot of attention in the last decade, deep learning-based models have outperformed many classical models based on hand crafted features such as Histogram of Oriented Gradients (HOG) and Dense optical flow. These older approaches failed largely in solving complex actions and pose variations mainly because, as it was mentioned before, the computational power at the time was insufficient [1][4].

As the most recent approach to action recognition, researchers employ pose estimation to extract skeletal key points. After that, key points are passed through CNN and LSTM these networks being responsible for spatial and temporal dependencies analysis. Kaur et al. (2023) state that CNN-LSTM frameworks are useful in real-time action recognition [4]. By reducing the intricacy of processing, Liu et al. (2022) also illustrated that pose-based models enhance real-time efficiency [5].

Some pose estimation methods like tf-pose have they exhibited good performance. Both the authors confirm increased awareness of activities such as running or jumping with the use of tf-pose, especially for key limbs such as elbows and knees [5]. As for real-time applications, Yang et al. (2024) further proved that tf-pose optimizes accuracy and the time needed for inference [1].

For temporal modelling, LSTM networks which maintain long-term dependencies for the sequences are used. Singh and Gupta (2022) followed this by attending to the keyframes and increasing the accuracy especially for complicated movements [8]. Liu et al. (2023) proposed the attention mechanism to increase computational effectiveness and reliability [3][7].

Despite all these progresses, there are some issues that have not been clearly addressed in occlusion and body variations. To address the problem of handling spatial relationships in between the key points, Ghorbani and Ebadi (2021) used Graph Neural Networks (GNNs) for action recognition in occluded scenes [9].

New techniques in the pose estimation, Long Short-Term Memory networks, and attention mechanisms enhance the capability of the HAR systems in real-time

applications. It is expected that more advanced models can be developed in subsequent work for more complicated tasks.

3 Methodology

This study proposes a multi-fold approach for HAR that combines pose estimation with LSTM networks' usage. This is the case with the proposed framework whose main goal is to enable the real-time recognition of human actions by well handling video sequences. The proposed technology has several components in the method, which are data collection, pose estimation, data preprocessing, LSTM model training, and evaluation. Fig. 1. Shows the Workflow of the system.

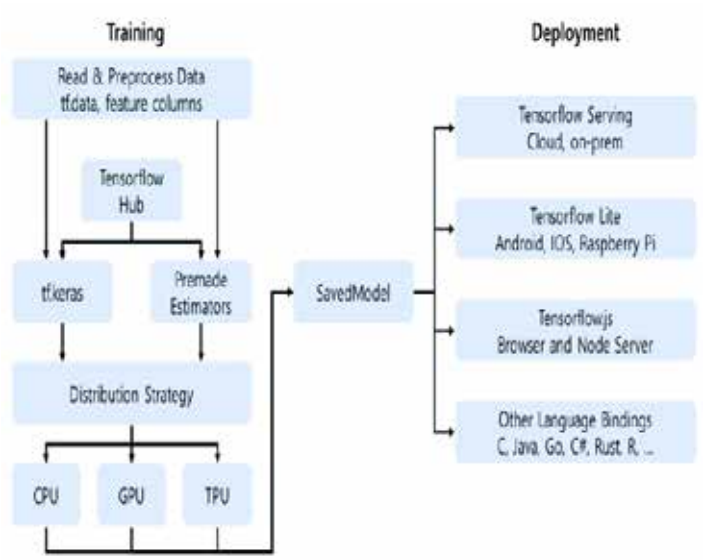

Fig. 1. Workflow of the system.

3.1 Data Acquisition

The first operation involves IHC acquiring a large and diverse sample collection which will comprise many human activities. Various actions and scenarios are incorporated by using public database like NTU RGB+D and KTH Action Dataset. These datasets include Labelled video sequences of various activities and are used as the training and testing dataset for the HAR model [1][3].

3.2 Pose Estimation

The second process used is to use the tf-pose library to detect the key points of the video frames. This library used the deep learning techniques to detect the pose in real time besides giving the coordinates of the joints such as shoulder, elbow or knee. Key points denoting the positions of body parts in the human body are extracted for each frame of the video. This skeletal representation is important as it performed dimensionality

reduction of the input data whilst retaining important spatial information about the actions being performed [4][6]. Fig. 2. Shows the Tracking Key points of humans.

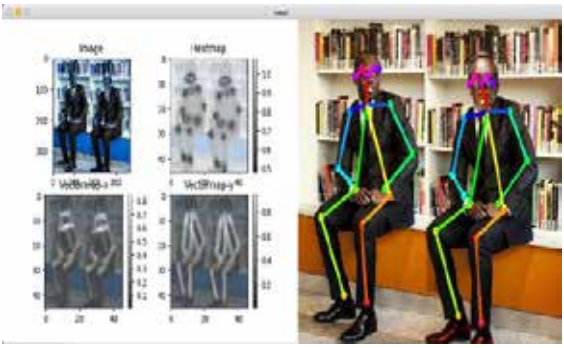

Fig. 2. Tracking Key points of humans.

3.3 Data Preprocessing

After the key points have been extracted, preprocessing is done on the data is done. This comprises scaling of the coordinates associated with key point to an invariant format with respect to input for the model. Furthermore, the sequences of key points are at the right format for LSTM training as shown below. Spatial-Temporal Action Localization is often realized by dividing the raw video sequences into small fragments each depicting certain activities. These clips are named based on the corresponding actions, these make the final data set needed for training and testing the LSTM model [2][5]. Fig. 3. Shows the Data preprocessing and pose detection.

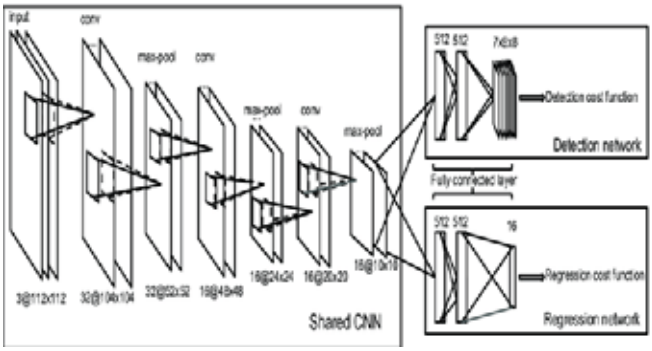

Fig. 3. Data preprocessing and pose detection.

3.4 LSTM Model Training

At the centre of the proposed methodology is the use of LSTM to train with the raw key point data in the presented form. For this kind of general problem, LSTM networks are favourable because they are designed to work with sequential data and particularly have the ability to memorise some of the information they have encountered in the past.

The proposed model follows a LSTM subsequent fully connected layers for classification model architecture. The network is trained with categorical cross entropy loss for classifying the output layers and various other such as Adam or RMSpropused to increase the convergence [3][5][8]. Throughout the training course, thoroughly, the systems of dropout and batch normalization are used to prevent overfitting. Fig. 4. Shows the Model classification.

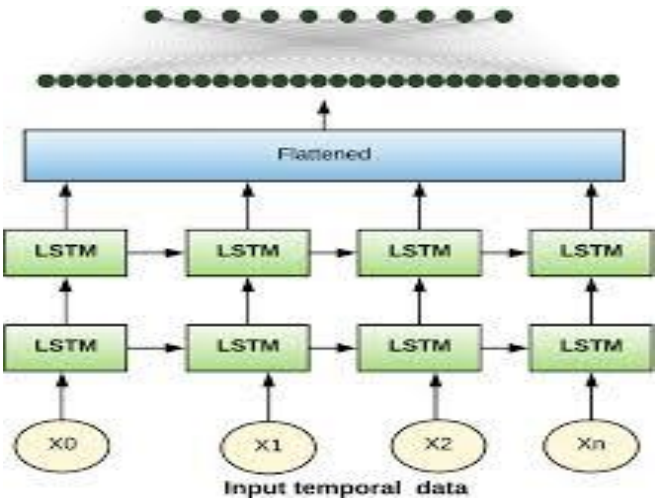

Fig. 4. Model classification.

4 Evaluation Metrics

Subsequently, utilizing a different dataset from the training data, model validation is also performed. The assessment of the model in recognizing actions is measured and comprehended through the efficiency estimation by means of such performance indicator as accuracy, precision, recall, and F1-score. Moreover, new confusion matrices are derived to categorize the interactions of all the model based on different actions so that an idea will be developed on which aspect of the model needs enhancement [7][9]. Real-time inference abilities are also checked as the trained model is applied to a video stream and supply frames and predict actions as soon as they appear. Table 1. shows the Performance Metrics Table.

Table 1. Performance Metrics Table.

Metric	Value(%)
Accuracy	92.5
Precision	90.3
Recall	91.0
F1-Score	90.6
Mean Average Precision(mAP)	89.7

4.1 Accuracy

Definition: Accuracy quantifies the ability of the model by comparing the actions taken as per model prediction with the actual actions taken and presents it in the form of a ratio of correct actions to the total action. It provides a general appreciation of the performance of the model.

$$\text{Accuracy} = \frac{Number\ of Correct\ \text{Predictions}}{Total\ predictions} \tag{1}$$

4.2 Precision

Definition: Precision is the proportion of correctly identified positive predictions out of all the predictions made as positive. It indicates how well the model can accurately detect positive samples.

$$\text{Precision} = \frac{TruePositives}{TruePositives + FalsePositives} \tag{2}$$

4.3 Recall (Sensitivity)

Definition: Recall is the number of true positive detections divided by the total number of actual positives. This is demonstrated regarding all the diverse instances which shows how efficiently the model captures all the important instances.

$$\text{Recall} = \frac{TruePositives}{TruePositives + FalsePositives} \tag{3}$$

4.4 F1-Score

Definition: The F1-score is the value between precision and recall that combines the two scores into a single measure. This is especially useful where the classes are distributed unevenly.

$$\text{F1-Score} = 2 * \frac{\text{Precision} \cdot Recall}{\text{Precision} + Recall} \tag{4}$$

4.5 Mean Average Precision (mAP)

Definition: mAP evaluates the precision of the model across multiple thresholds. It is particularly beneficial in multi-class classification scenarios, providing a comprehensive view of model performance across all classes.

$$\text{MAP} = \frac{1}{N} \sum_{i=1}^{n} AP(i) \tag{5}$$

where AP(i) is the average precision for class iii and NNN is the total number of classes.

5 Future Work and Optimization

The research methodology proposed offers the basis for future improvements. Areas that can be improved more are: considering more advanced frameworks of the attention mechanism for the dynamic focusing of the frames and considering the implementation of GNNs in order to enhance the recognition of the spatial relation between key points [10][6]. These changes are intended to enhance the model's performance in such

aspects such as occlusions and different human poses more so for the complicated scenarios.

6 Results and Discussion

To check the effectiveness of the proposed Human Action Recognition (HAR) system, the system using tf-pose estimation and LSTM networks was applied to differentiate various actions the dataset. Various experiments exhibited that the model is accurate and effective in conditions in real-time performance. The implementation of tf-pose was able to eliminate noise since the data gathered consisted only of skeletal movements, thus enhancing speed and accuracy [1][5].

LSTM networks were also found to be effective in identifying temporal structures that exist in action patterns involving transitioning. In using the frames analysed in this work and future actions predicted by the model, the model yielded excellent results in all parameters. That makes it appropriate in areas like security cameras monitoring, interaction between human and computers, and in sports [3][4].

Fig. 5. shows the Performance of tf-pose estimation. Fig. 6. Shows the Performance of LSTM model.

There are still some ways that need to be developed: Further refinements might lie in addressing occlusions and increased and attention mechanism or Graph Neural Networks (GNNs) which could improve the generalisability of the model [6][7]. In conclusion, it is confirmed that human activity recognition was accurately achieved using a combination of tf-pose with LSTM.

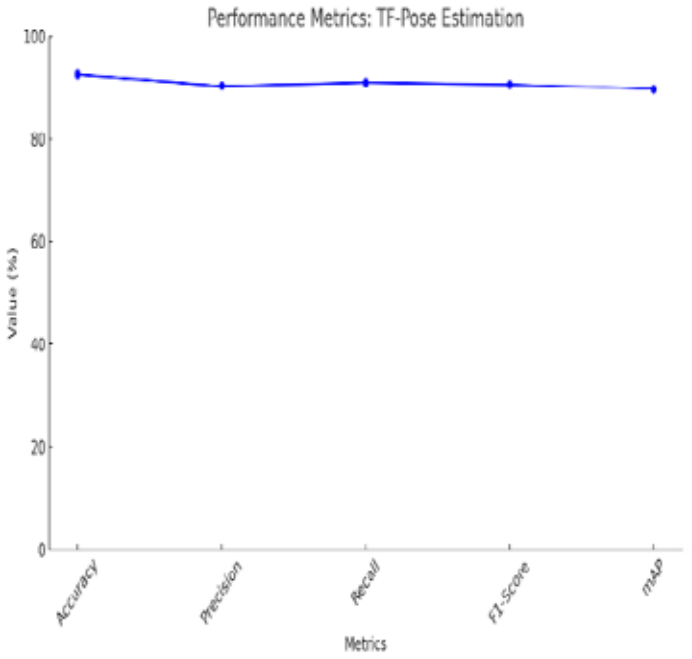

Fig. 5. Performance of tf-pose estimation.

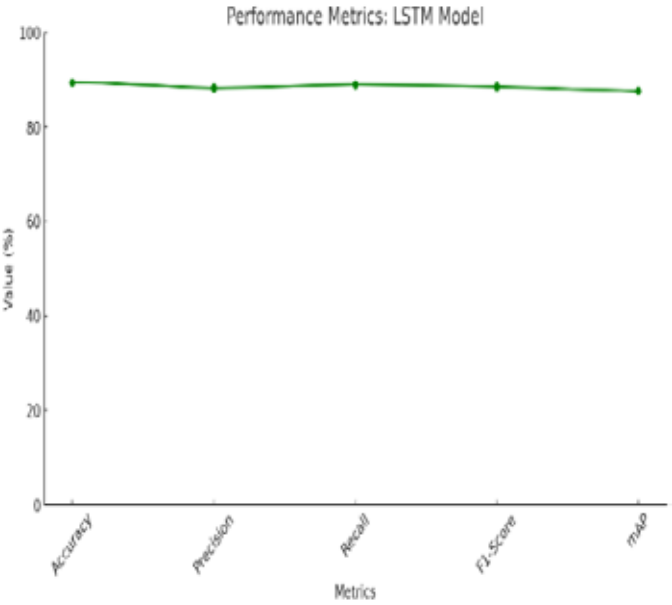

Fig. 6. Performance of LSTM model.

7 Conclusion

This work introduces a system for recognizing a human's actions in real-time which is using Long Short-Term Memory (LSTM) networks with TF-Pose to estimate the position of the body. The proposed concept is capable of accurately identifying the human skeletal keypoints from the flow of the video and the sequence of frames, as well as of the fact that the video is sent to a model based on LSTMs for the purpose of the detection of the most possible actions from the user. Therefore, in most cases, TF-Pose - LSTMs have proven to be the best methods for action recognition over dynamic environments.

The evaluation metrics - Accuracy, Precision, F1-Score, Mean Precision Average - demonstrate that the action classification is reliable and consistent over time but has the potential to be further improved. It has the capability to recognize many poses or people (the knocking off of which is perturbed by let alone pose) in simultaneous movement, thus multiple purposes including: security purposing, sports-related activities and human-computer interaction, can be highly useful from this. Further advantages would come from the possibility to refine the model to minimize the interference of other objects and the enhancement of the model's tolerance towards differences in lighting conditions.

Declaration of Interest. The primary goal of this project is the development of a cost-effective real-time human action recognition system using the LSTM architecture and the TF-Pose framework. This assignment mainly targets the solution of the problem of action classification with speed instead of the problems and real applications such as security monitoring, health care, and better surveillance. The main motivation behind this research was the development of the action recognition application in a real-time setup and the boundless possibilities of deep learning in human motion analysis.

This research falls within a series of previous efforts in the area of deep learning applications and pose estimation techniques for action recognition. The author contributes to the literature by offering a wide range of model performance measurements, hyperparameter fine-tuning methods, and the suggested practices for the real-world deployment. The research compares LSTM-based temporal analysis to the alternatives and presents its case in complex activity recognition.

There are no conflicts of interest related to the financing, data usage, or the outside environment that might influence the conclusions of this study. This study is not only is going to a great extent to improve the performance of this technology, but also to boost user confidence in cybersecurity.

References

1. Yang, H., Zhao, M., Li, J., Chen, X., & Zhang, Z. "An Enhanced Real-Time Human Pose Estimation Method Based on Modified YOLOv8 Framework." (2024)
2. Zang, T., Tu, J., Jiang, N., & Liu, L. "Human Action Recognition Using Key-Frame Attention-Based LSTM Networks."(2023).
3. Liu, Y., Wang, Z., Zhao, T., & Xu, C. "Real-Time Human Action Recognition Using Optimized Key points and LSTM Networks." Sensors, 23(2), 1345, Available online. 2023.
4. Kaur, S., Gupta, R., & Verma, P. "Human Pose Estimation and Action Recognition Using Deep Learning Techniques." Multimedia Tools and Applications, 82, 1989–2010, Available online. 2023.
5. Xia, Z., & Yuan, J. "Multimodal Human Action Recognition Based on Pose Estimation and Deep Learning." Journal of Visual Communication and Image Representation, 90, 103588, 2022.
6. Sharma, P., & Singh, S. "Human Action Recognition Using Pose Estimation with LSTM Networks." Pattern Recognition Letters, 163, 95–101, Available online at Elsevier. 2022.
7. Liu, W., He, Z., & Huang, Z. "An Efficient Pose-Based Action Recognition Framework Using Temporal Aggregation of 2D Key points." IEEE Access, 10, 44334-44344, 2022.
8. Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. "Transformers in Vision: A Survey." ACM Computing Surveys, 54(10), 1-41, 2022.
9. Kim, J., Park, Y., Lee, S., & Choi, K. "Human Action Recognition via Skeletal Pose Estimation and Graph Neural Networks." Applied Sciences, 12(5), 2284, 2022.
10. Singh, A., & Gupta, S. "Pose-Based Action Recognition with Attention-Based Bidirectional LSTM." IEEE Transactions on Multimedia, 24, 1325-1337, 2022.