A Comparative Study on Deep Learning and Machine Learning Models for Human Action Recognition in Aerial Videos

Surbhi Kapoor
UIET, Panjab University, India
surbhi31892@gmail.com

Akashdeep Sharma
UIET, Panjab University, India
akashdeep@pu.ac.in

Amandeep Verma
UIET, Panjab University, India
amandeepverma@pu.ac.in

Vishal Dhull
UIET, Panjab University, India
mdhull07@gmail.com

Chahat Goyal
UIET, Panjab University, India
chahatgoyal1999@gmail.com

Abstract: *Unmanned Aerial Vehicle* (*UAV*) *finds its significant application in video surveillance due to its low cost, high portability and fast-mobility. In this paper, the proposed approach focuses on recognizing the human activity in aerial video sequences through various keypoints detected on the human body via OpenPose. The detected keypoints are passed onto machine learning and deep learning classifiers for classifying the human actions. Experimental results demonstrate that multilayer perceptron and SVM outperformed all the other classifiers by reporting an accuracy of 87.80% and 87.77% respectively whereas LSTM did not produce very good results as compared to other classifiers. Stacked Long Short-Term Memory networks (LSTM*) *produced an accuracy of 71.30% and Bidirectional LSTM yielded an accuracy of 76.04%. The results also indicate that machine learning models performed better than deep learning models. The major reason for this finding is the lesser availability of data and the deep learning models being data hungry models require a large amount of data to work upon. The paper also analyses the failure cases of OpenPose by testing the system on aerial videos captured by a drone flying at a higher altitude. This work provides a baseline for validating machine learning classifiers and deep learning classifiers against recognition of human action from aerial videos.*

Keywords: *Human action recognition, openpose, unmanned aerial vehicle, kNN, decision trees, random forests, SVM, multilayer perceptron, LSTM.*

Received October 29, 2021; accepted October 26, 2022
https://doi.org/10.34028/iajit/20/4/2

1. Introduction

Unmanned Aerial Vehicle **(**UAVs) have revolutionized the aerial world due to their numerous benefactors like ease of deployment, low maintenance cost, high portability, high-mobility and ability to hover at any point. Nowadays, drones are ubiquitous and are actively being used in several applications such as sports, entertainment, agriculture, forest monitoring, military and surveillance. Human action recognition is one of the prominent applications of UAVs and is very less explored by the researchers due to the complex nature of human actions.

Human actions can be categorized into gestures, atomic actions, human to human interactions, human behavior and event actions and their recognition is a daunting task. Human pose estimation refers to the problem of estimating the joints on the human body. Taking human pose estimation into account, relevant datasets have been published namely Frames Labeled In Cinema (FLIC), Leeds Sports Pose (LSP), Common Objects in Context (COCO) and AI Challenger-Human Keypoint Detection (AIC-HKD) for images and Penn Action (University of Pennsylvania), Joint-annotated Human Motion Data Base (J-HMBD) and PoseTrack for videos. These datasets are captured either by stationary cameras or sequences from movies or YouTube videos and do not involve camera movement. Some progress was made in the domain of human activity recognition, considering the above mentioned problems by introducing datasets such as HMDB-51 and UCF-101. But the problem of recognizing the human actions from an aerial video remains an unsolved problem. Datasets namely Drone-Action [8], UAV-Gesture [7] addressed the problem of estimating the human pose in aerial videos by estimating the joints (keypoints) on the human body and forming a skeleton by joining those keypoints. The detailed discussion on these datasets and techniques is presented in section 2.

This paper is based on estimating the human pose and recognizing the actions based on the keypoints. The technique used in extracting the keypoints is OpenPose [3]. It is an open source library developed for detecting 2D poses in images and videos. It detects

keypoints on the human body, foot, face and hand. It can produce a total of 135 keypoints on a single frame. Once the keypoints are extracted, they can be processed and passed to various classifiers for classifying the human actions. In this work, five different machine learning classifiers have been selected to perform classification on human pose keypoints i.e., k-Nearest Neighbors (kNN), decision trees, random forests, Support Vector Machine (SVM), and multi-layer perceptron along with two variants of Long Short-Term Memory networks *Long* Short-Term Memory Networks (LSTM) i.e., Stacked LSTM and Bidirectional LSTM.

The main contributions made in this study are summarized as follows:

1. Comparing and analyzing performance of various machine learning classifiers for human action recognition.
2. Classifying human actions using LSTM with OpenPose as a baseline for feature extraction.

The rest of the paper is organized as follows. Section 2 covers the related work. Section 3 discusses the problem statement of the work. Section 4 contains the proposed work followed by the description of the dataset in section 5. Experimental results are shown in section 6. Section 7 discusses the failure cases of OpenPose and section 8 concludes the paper along with the future work.

2. Related Work

With the advent of deep learning, Convolutional Neural Network (CNN) came into picture, marking a prominent impact in the field of human pose estimation. State of the art methods [5, 10, 14, 15] in human pose estimation have used CNN as the backbone to improve the performance of the techniques. Zebhi *et al.* [16] and Pushparaj and Arumugam [11] have used spatio temporal modules for recognizing human actions from surveillance videos. Human pose estimation in aerial imagery is still an emerging area of research. Drones can be used to identify the people lying on the ground. Andriluka *et al.* [1] evaluated the different traditional approaches namely histogram of oriented gradients, poselets, Deformable Parts Model (DPM) in order to identify the persons (victims) lying on the ground probably due to some injury. The experiments were conducted on the images collected by a UAV in an indoor office environment simulating the conditions of self occlusion and partial visibility of body parts arising in search and rescue operation. Some studies have focused on human motion analysis and gait estimation using a drone. Perera *et al.* [9] used dynamic classifier for pose estimation and estimated the trajectory of the human by exploiting 3D skeletons. Penmetsa *et al.* [6] estimated the human pose by using pictorial structures and classified the estimated pose as suspicious by exploiting the orientation of the limbs. Inspired by [6], Singh *et al.* [12] identified the violent actions by replacing the pictorial structures with their proposed, ScatterNet Hybrid Deep Learning (SHDL) network in order to estimate the human pose followed by classification of the pose by SVM based on orientation of the limbs. The authors did not make the dataset publicly available. To address the scarcity of dataset, Perera *et al.* [8] presented a new dataset, Drone-Action for action recognition in aerial videos. The dataset was recorded by a drone hovering at a low altitude and was evaluated using Pose Based Convolutional Neural Network (P-CNN) and High Level Pose Features (HLPF) descriptors.

Human activity recognition is a very wide domain and gesture recognition is a sub module of human action recognition which highly depends on the joints extracted from the human body. Researchers have used human pose features to identify gestures in aerial imagery. One of the worth mentioning datasets studied in this field is UAV-Gesture [7] captured by a hovering drone. The authors identified the human pose by estimating the body joints using OpenPose and evaluated their dataset by P-CNN descriptor. Another dataset, given by Song *et al.* [13] containing 24 actions, was introduced for aircraft handling. Bolin *et al.* [2] used a pre-trained CNN for joint estimation and applied finite state machine to select the UAV's action.

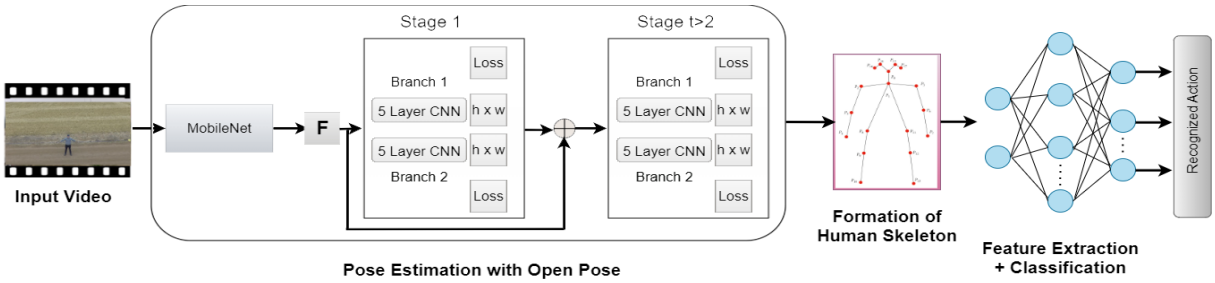


Figure 1. Workflow of the proposed approach.

Chen *et al.* [4] combined the action recognition with gesture recognition module to control the drone flight actions by using spatio temporal Graph Convolutional Network (GCN) to recognize the human actions. As cited in the above literature, the task of human action recognition in aerial videos has not gained much

success. The current study focuses on UAV captured videos in which human pose extraction serves as a base for identification of human action recognition.

3. Problem Statement

Human action recognition is an active research topic in the field of computer vision and image processing and finds its application in video surveillance, behavior analysis, health care systems etc., The problems contributing to the slow progress of this domain includes the lack of proper aerial datasets as most of the human action recognition studies are based on fixed cameras. Another reason adding to the slow progress is complexity of human actions. The human actions are so complex in nature that it is very difficult to model them. The state of the art algorithms are fallible as they do not report much accuracy. The proposed work aims to recognize the human actions based on human skeleton data and provides a comparison between machine learning and deep learning models.

4. Proposed Work

This paper aims to recognize human actions from a video sequence captured from drones. In the proposed approach, action recognition is based on the keypoints, extracted by OpenPose, which are joined together to form a human skeleton. The combination of these keypoints serves as the features which are fed to the machine learning classifiers as input for recognizing the actions. The overall workflow of the process is shown in Figure 1. We will briefly describe each module in the upcoming section.

4.1. Pose Estimation

OpenPose is used to extract the keypoints on human body. It is built upon convolutional neural networks and it is the first open source system which can jointly localize keypoints on human body, hands, foot and face. The model can detect a total of 135 keypoints in a single frame. We are focused on mainly 18 keypoints as we keep our approach limited to human body only because the other parts of the body like hand and foot do not play a significant role in activity recognition in aerial videos. OpenPose outputs 18 joints including 5 joints on head i.e., 1 for head, 2 for eyes and 2 for ears. Each joint is characterized by x and y coordinates, summing up to 36 values for each skeleton. OpenPose is originally trained on MS COCO and MPII datasets.

4.2. Preprocessing the Skeleton Data and Feature Extraction

In our action recognition task, facial key points do not play a vital role. In order to reduce the complexity of the problem, we have eliminated these 5 key points from each skeleton. Moreover, there can be missing value of some joints in few frames due to self occlusion. Rather than discarding those frames, the missing value of the joint is filled according to its relative value in the previous frame. The processed raw skeleton data is used to extract features for classification. We have considered three features *i.e.* velocity of the body, velocity of joints and normalized joint positions.

4.3. Classification

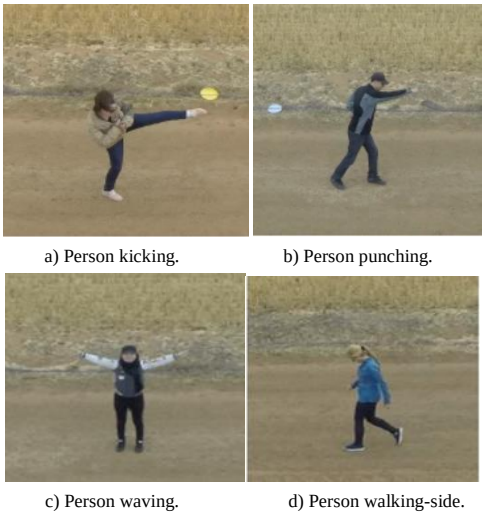
Actions are classified by feeding the features into five machine learning classifiers namely SVM, KNN, Decision trees, Random forests, Multi layer perceptron and two variants of deep learning i.e., stacked LSTM and bidirectional LSTM.

5. Dataset

The proposed approach is evaluated on Drone-Action dataset [8]. The dataset is captured by a drone hovering at a low altitude (8-12m) in a wheat field with a human performing the action in the centre. There are a total 240 HD videos having a resolution of 1920x1080 collected at 25fps. A total of 13 actions were recognized namely ‘Punching’, ‘Clapping’, ‘Hitting’, ‘Bottle’, ‘Hitting Stick’, ‘Jogging front back’, ‘Jogging side’, ‘Kicking’, ‘Running front back’, ‘Running Side’, ‘Stabbing’, ‘Walking front back’, ‘Walking Side’ and ‘Waving Hands’. A glimpse of the dataset is shown in Figure 2.

6. Experimental Results

A total of 13 actions were classified by different machine learning algorithms and two variants of LSTM. The hyperparameters used in all the classifiers are explained below.



a) Person kicking. b) Person punching.

c) Person waving. d) Person walking-side.

Figure 2. Glimpse of the drone-action dataset [8].

Mobilenet thin is used as the base network for OpenPose. Various hyperparameters are tuned in order

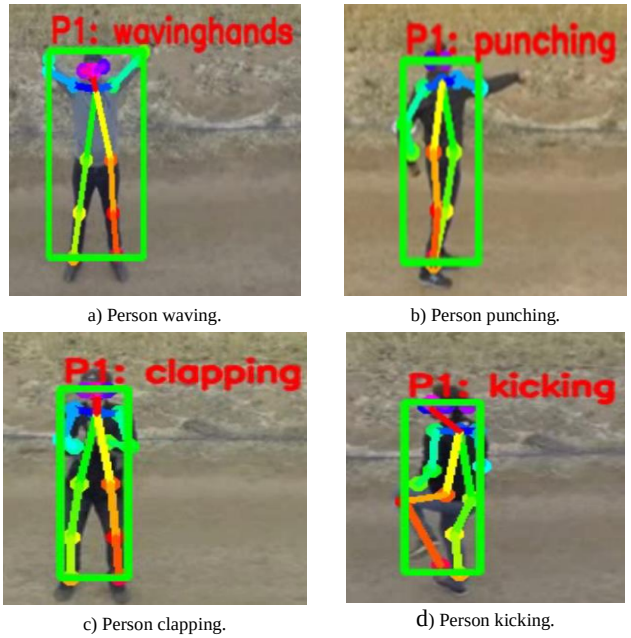
to measure the performance of the classifiers. In KNN, choosing an optimum value of K is necessary which can turn the results in one's favor. Too small value of K will make the results blow by local neighborhood and with a larger value of K, the results will depend on the most likely response in the data. The experiments were performed by taking the number of neighbours as 5 giving an accuracy of 82.56%. For SVM, different kernels were tried to get the optimum result. The best results were produced by rbf kernel, setting gamma to scale and reguralization parameter as 1, yielding an overall accuracy of 87.77%. In MLP, we have experimented with combinations of activation functions and solvers. The optimum results were reported with tanh activation function and adam optimizer with 3 hidden layers having the number of neurons as 40, 30, and 20 respectively and performing 1000 iterations. The model converged at 87.80% accuracy. The Random forests do not require much of the hyperparameter. Max_depth in random forests and decision tree is set to 10 after experimenting at various depths and leaving all the other parameters to default value.

The evaluation metric selected for the experiments were precision, recall and accuracy. We report the highest accuracy of 87.80% on test data achieved by multilayer perceptron and lowest accuracy of 64.25% achieved by decision tree. The average accuracy obtained on test set by all the classifiers is 79.3%. In the category punching, multi layer perceptron and SVM gave the same results whereas decision trees produced the worst results by yielding a precision of 0.50. The category running_fb, yielded very low precision values with all the classifiers. In Random forests, the precision value in the category of running_fb is zero which means random forests were not able to classify even a single example of this category. MLP performed very well for all the other categories except running_fb. The moderate set of values were observed by KNN. The architecture of stacked LSTM used in this paper comprises of two layers of LSTM with 34 hidden cells each along with tanh activation function. The choice of optimizer is adam optimizer with a learning rate of 0.0025 and softmax cross entropy loss.

The model is trained for 80 epochs with a batch size of 128 achieving an accuracy of 71.30%. Several experiments were performed with varying hyperparameters but the hyperparameters listed above present the best results for the given task as the model started converging. The model of bidirectional LSTM used here consists of two forward layers and two backward layers with 34 hidden cells and tanh activation function as used in stacked LSTM. All the other hyperparameters are same as with stacked lstm. The accuracy yielded by bidirectional LSTM is 76.04%. The action classification output of mlp classifier is shown in Figure 3 and the results for precision and recall of all the classifiers are presented in Figures (4-10). The accuracy obtained by all the classifiers is given in Table 1.

Table 1. Accuracy obtained by different classifiers.

Classifier	Accuracy
Multilayer perceptron	87.80%
SVM	87.77%
KNN	82.56%
Random forests	74.32%
Decision trees	64.25%
Stacked LSTM	71.30%
Bidirectional LSTM	76.04%



a) Person waving. b) Person punching.

c) Person clapping. d) Person kicking.

Figure 3. Action classification output of mlp classifier.

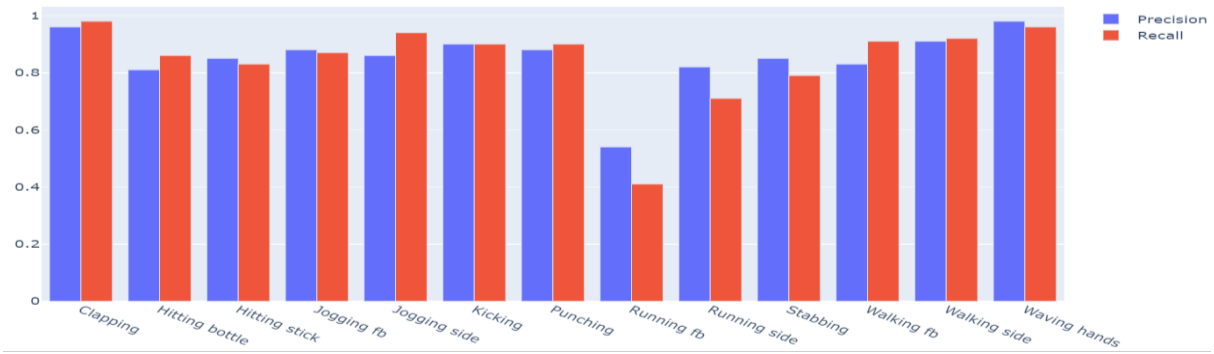


Figure 4. Results obtained after classification by multilayer perceptron.

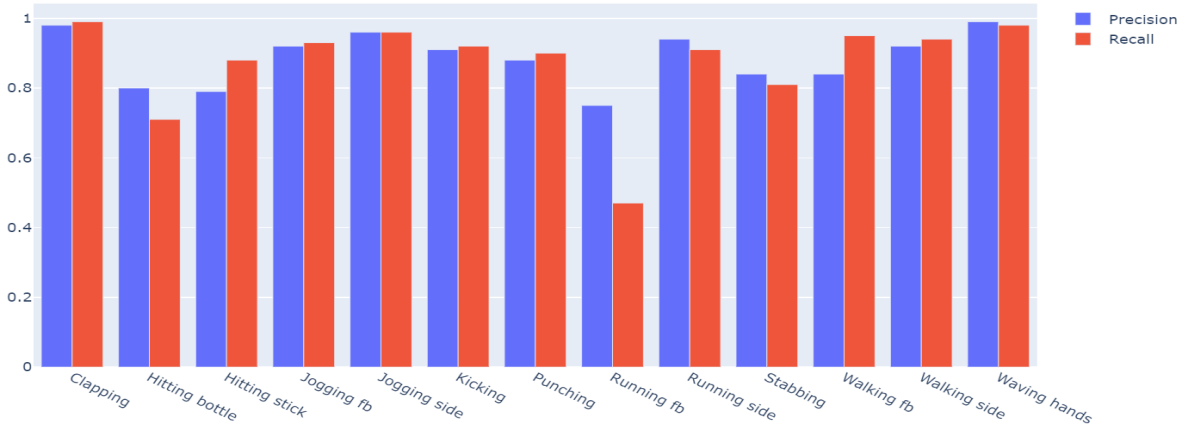


Figure 5. Results obtained after classification by SVM.

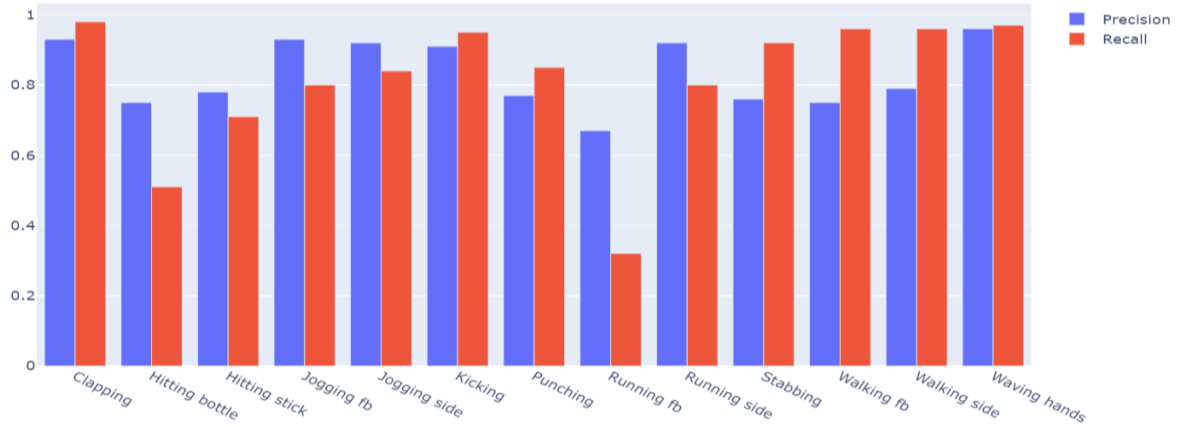


Figure 6. Results obtained after classification by KNN.

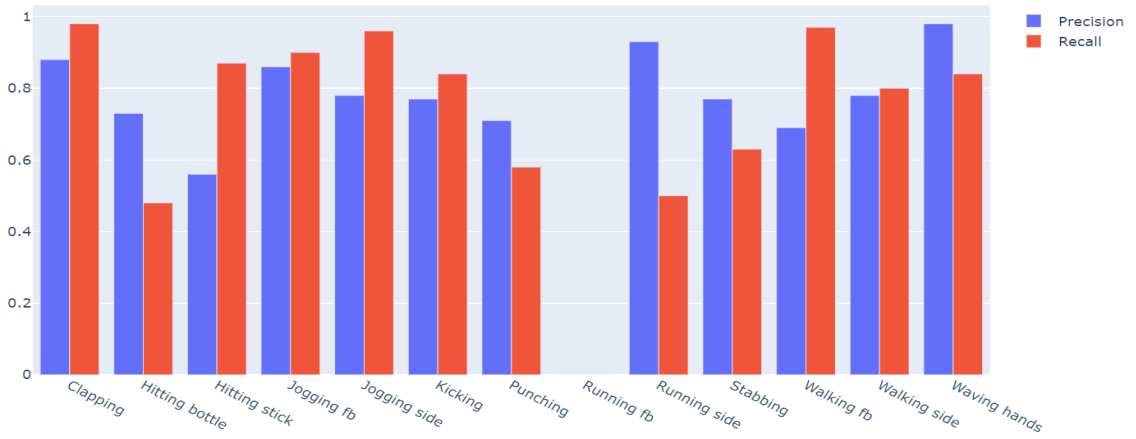


Figure 7. Results obtained after classification by random forests.

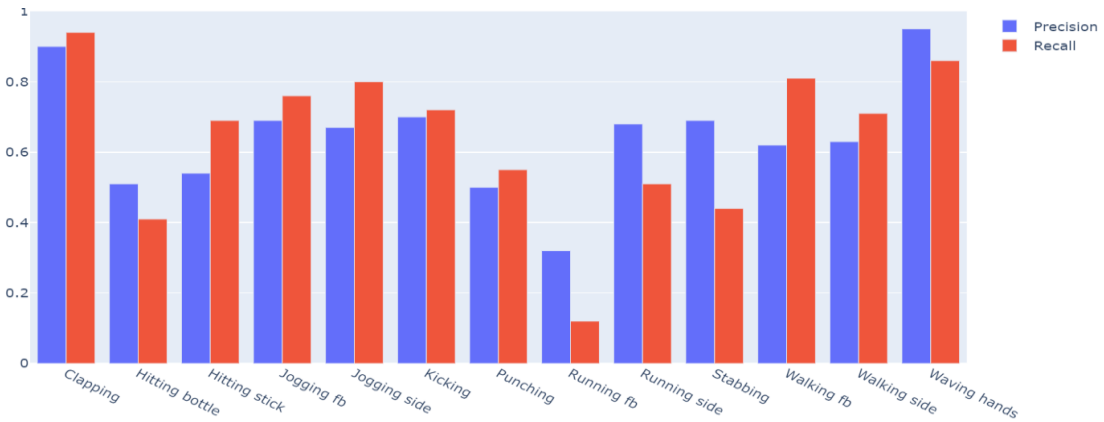


Figure 8. Results obtained after classification by decision Tree.

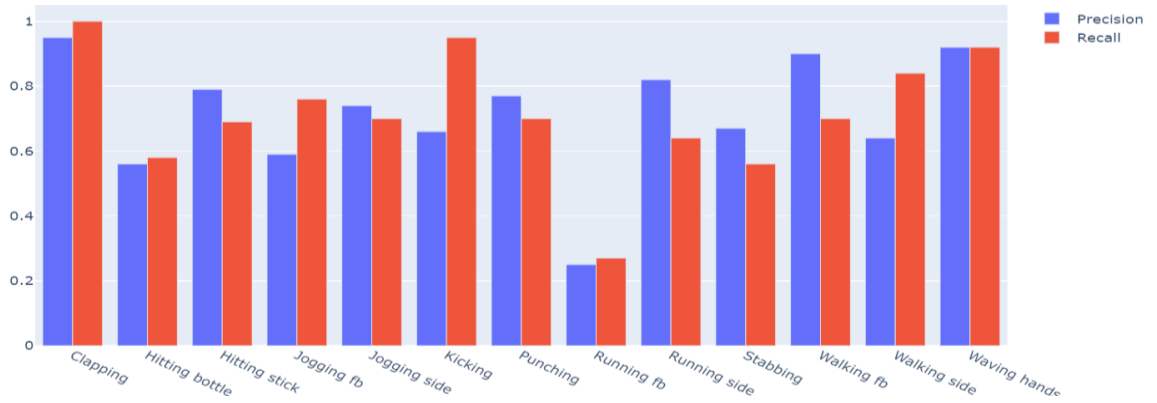


Figure 9. Results obtained after classification by stacked LSTM.

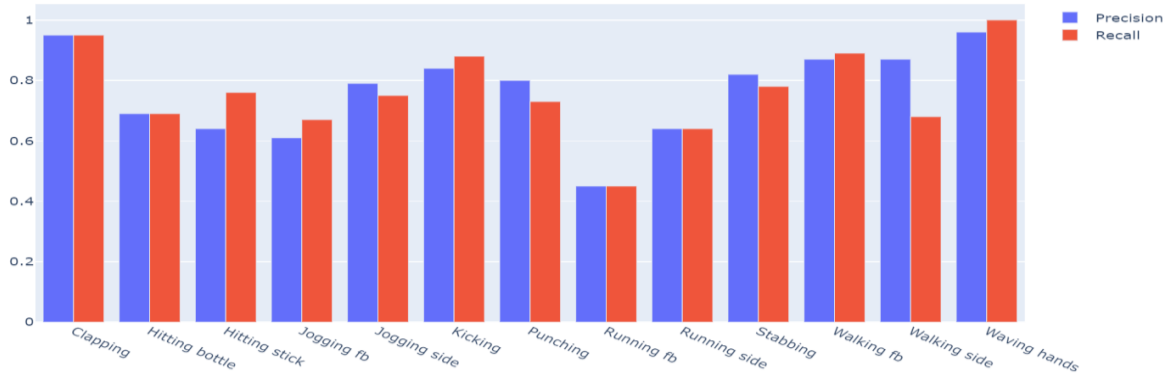


Figure 10. Results obtained after classification by bidirectional LSTM.

7. Analysis of OpenPose Failure Cases

OpenPose is able to detect the keypoints well in aerial images captured at a lower altitude but not at higher altitude. The paper also experimented with OpenPose on the aerial images captured by a drone flying at a higher altitude (15-20m) approx in an outdoor setting. The system was not able to localize keypoints properly on the human body, rather the keypoints were located somewhere else in the image. Figure 11 shows the failure of OpenPose. The possible solution to this problem is fine tuning the OpenPose with images taken at a higher altitude.

Figure 11. Failure case of OpenPose.

8. Conclusions

This study has presented a system for action recognition system based on 2D human skeleton data using UAV captured videos. Firstly, OpenPose was used to estimate 2D human skeleton data. The extracted information from human skeleton was fed to five different classifiers for action classification. The experiments are conducted on Drone-Action dataset containing 13 actions. Experiments show that multilayer perceptron outperform all other classifiers yielding the highest average accuracy of 87.8% whereas lowest accuracy of 64.25% is obtained by decision trees. Clapping and waving are the most correctly recognizable actions which have produced promising results with all the classifiers. MLP and SVM have shown better results than deep learning models due to the lesser availability of data which is the major reason behind the low performance of LSTM models. In future, we aim to recognize more complex activities involving more numbers of persons, expanded dataset and fine tuning the OpenPose system for aerial images captured at higher altitude.

References

[1] Andriluka M., Schnitzspan P., Meyer J., Kohlbrecher S., Petersen K., Stryk O., Roth S., and Schiele B., "Vision Based Victim Detection from Unmanned Aerial Vehicles," *in Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, Taipei, pp. 1740-1747, 2010. DOI: 10.1109/IROS.2010.5649223

[2] Bolin J., Crawford C., Macke W., Hoffman J., Beckmann S., and Sen S., "Gesture-Based Control of Autonomous Uavs," *in Proceedings of the 16th Conference on Autonomous Agents and*

MultiAgent Systems, Brazil, pp. 1484-1486, 2017. https://dl.acm.org/doi/10.5555/3091125.3091337

[3] Cao Z., Hidalgo G., Simon T., Wei S., and Sheikh Y., "Openpose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 172-186, 2019. https://doi.org/10.1109/TPAMI.2019.2929257

[4] Chen B., Hua C., Li D., He Y., and Han J., "Intelligent Human-UAV Interaction System with Joint Cross-Validation Over Action-Gesture Recognition And Scene Understanding," *Applied Sciences*, vol. 9, no. 16, pp. 3277, 2019. https://doi.org/10.3390/app9163277

[5] Li M., Zhou Z., Li J., and Liu X., "Bottom-up Pose Estimation of Multiple Person with Bounding Box Constraint," *in Proceedings of 24th International Conference on Pattern Recognition*, Beijing, pp. 115-120, 2018. https://doi.org/10.48550/arXiv.1807.09972

[6] Penmetsa S., Minhuj F., Singh A., and Omkar S., "Autonomous UAV for Suspicious Action Detection Using Pictorial Human Pose Estimation and Classification," *ELCVIA: Electronic Letters on Computer Vision and Image Analysis*, vol. 13, no. 1, pp. 18-32, 2014. DOI: 10.5565/rev/elcvia.582

[7] Perera A., Law Y., and Chahl J., "UAV-GESTURE: A Dataset for UAV Control and Gesture Recognition," *in Proceedings of the European Conference on Computer Vision Workshops*, 2018.

[8] Perera A., Law Y., and Chahl J., "Drone-Action: An Outdoor Recorded Drone Video Dataset for Action Recognition," *Drones*, vol. 3, no. 4, pp. 82, 2019. https://doi.org/10.3390/drones3040082

[9] Perera A., Law Y., and Chahl J., "Human Pose and Path Estimation from Aerial Video Using Dynamic Classifier Selection," *Cognitive Computation*, vol. 10, no. 6, pp. 1019-1041, 2018. https://doi.org/10.1007/s12559-018-9577-6

[10] Pishchulin L., Insafutdinov E., Tang S., Andres B., Andriluka M., Gehler P., and Schiele B., "Deepcut: Joint Subset Partition and Labeling for Multi Person Pose Estimation," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4929-4937, 2016. https://doi.org/10.48550/arXiv.1511.06645

[11] Pushparaj S. and Arumugam S., "Using 3D Convolutional Neural Network in Surveillance Videos For Recognizing Human Actions," *The International Arab Journal of Information Technology*, vol. 15, no. 4, pp. 693-700, 2018. https://www.iajit.org/PDF/July%202018,%20No.%204/8768.pdf

[12] Singh A., Patil D., and Omkar S., "Eye in the Sky: Real-Time Drone Surveillance System (DSS) for Violent Individuals Identification Using Scatternet Hybrid Deep Learning Network," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Salt Lake, pp. 1629-1637, 2018. DOI: 10.1109/CVPRW.2018.00214

[13] Song Y., Demirdjian D., and Davis R., "Tracking Body and Hands for Gesture Recognition: NATOPS Aircraft Handling Signals Database," *in Proceedings of Face and Gesture*, Santa Barbara, pp. 500-506, 2011. DOI: 10.1109/FG.2011.5771448

[14] Toshev A. and Szegedy C., "Deeppose: Human Pose Estimation Via Deep Neural Networks," *in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Columbus, pp. 1653-1660, 2014. https://doi.org/10.1109/CVPR.2014.214

[15] Wei S., Ramakrishna V., Kanade T., and Sheikh Y., "Convolutional Pose Machines," *in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 4724-4732, 2016.

[16] Zebhi S., Almodarresi S., and Abootalebi V., "Human Activity Recognition Based on Transfer Learning with Spatio-Temporal Representations," *The International Arab Journal of Information Technology*, vol. 18, no. 6, pp. 839-845, 2021. https://doi.org/10.34028/iajit/18/6/11

Surbhi Kapoor has completed her B.Tech from GNDEC,Punjab, India in 2014 and M.Tech in 2016 She is pursuing her Ph.D in Computer Science and Engineering from UIET, Chandigarh, India. Her research interests include object detection, image and video analytics.

Akashdeep Sharma is currently working as an assistant professor in Computer Science and Engineering at UIET, PanjabUniversity, Chandigarh, India. His research interests include video analytics, object detection and tracking and classification.

Amandeep Verma is currently working as an associate professor in Information and Technology at UIET, PanjabUniversity, Chandigarh, India. Her area of interest includes parallel and distributed computing, computer networks and IoT.

Vishal Dhull is undergraduate in 3rd year of Bachelor's degree in Computer Science andEngineering at UIET, PanjabUniversity, Chandigarh.

Chahat Goyal is an undergraduate in 4th year of Bachelor's degree in Computer Science andEngineering at UIET, PanjabUniversity, Chandigarh.