**HUMAN ACTIVITY CLASSIFICATION FROM DRONE VIDEOS**

Anuj Rajodiya Damerla Rampavan

IIT (ISM) Dhanbad Bennett University

[anujrajodiya@gmail.com](mailto:anujrajodiya@gmail.com) [dr9236@bennett.edu.in](mailto:dr9236@bennett.edu.in)

M. Krishna Pranathi Kaleel Rahman

SRK Institute of technology Sri Ramakrishna Engineering College

[pranathi.meegada@gmail.com](mailto:pranathi.meegada@gmail.com) [rkaleel375@gmail.com](mailto:rkaleel375@gmail.com)

**Mentors**

Dr.K. Biswas sir

Anugrah Srivastava sir

1. **Abstract**

We contemplate the machine-driven recognition and classification of human actions from drone videos. In our everyday life surveillance has become crucial for secured life which can be accomplished efficiently by activity classification from drone videos. Drones square measure presently being employed for industrial, public safety, and research project functions. Such human activity recognition is predicated on information gathered from videos. It's wide applications are police work, video categorization, biometrics and human-computer interaction. This would be a good application serving different purposes. For classification we used random forest, a machine learning model and boosting technique in it. Finally, different activities from drone videos are classified accordingly using this model.

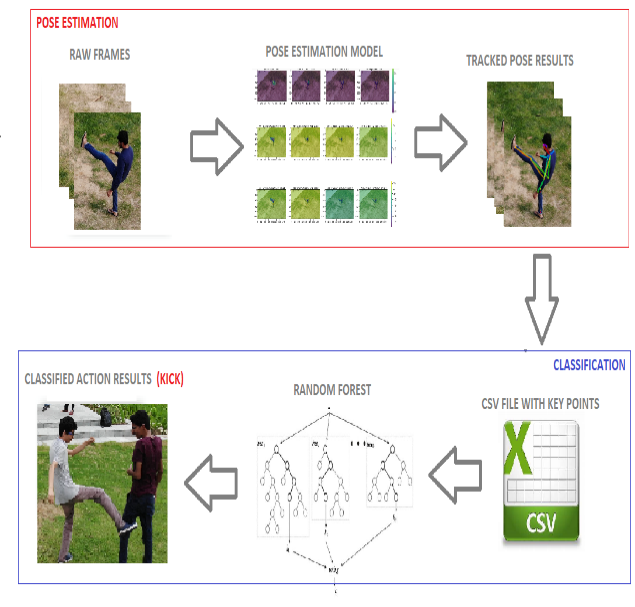
1. **Introduction**

Human action is that the real-world depiction of the person’s intentions, behaviour and thoughts. Classifying human actions finds applications in a very style of domains like video surveillance, behaviour analysis etc. The need for human activity classification is that if we train machines to automatically do this then any unusual

activity can be identified immediately from surveillance cameras and corresponding action can be initiated within no time, so that security can be improved and simplified. This includes human recognition, pose estimation from key point abstraction and finally action classification. But, accurate recognition of actions may be a difficult task. We should take into consideration the circumstances beneath that the video was taken, environment, viewpoint variations and every one since, in this totally different action categories and motion patterns could occur.

The classification of human activities includes classifying the actions as abnormal or traditional. This is done by initial estimation of pose by finding the key points then, classifying the estimated pose into one in all that binary classification mentioned above. Human body is very articulated with several degrees of freedom to support advanced movements. making a comprehensive model which will represent these movements exploitation proof collected from pictures is so terribly difficult, since it poses a retardant in an exceedingly very high dimensional area.

We can take this information of videos from static camera additionally since it should not cowl all the angles and to enhance the performance we tend to use drones for this purpose. After this all the data collected is pre-processed and model is applied. After training the model, it'll be capable of classifying actions from drone videos.



In the paper, we tend to develop a model for action recognition, since classical approaches to the matter involve training machine learning models, like ensembles of decision trees. In this we tend to use boosting technique. To avoid over-fitting in random forest, we optimized a calibration parameter that governs the quantity of options that square measure indiscriminately chosen to grow every tree from the bootstrapped knowledge, reduced variance. Our dataset consists of three hundred 3D sequences of 14 activities performed by single additionally as multiple persons in varied views. Our methodology is period. To improve the accuracy of our model we tend to perform adding additional data, treat missing and outlier values, parameter calibration etc. From this, we determine traditional and abnormal activities of humans from drone videos.

**3. Related work**

Many others had worked on this activity classification before. There are many surveys within the act recognition and classification fields. It originate to 1999. In an analysis paper ramnathan clearly mentioned regarding the challenges in act recognition from drone videos. Work on pose estimation is finished by wang that became useful in later works including pose estimation of humans. Characterization of act recognition strategies into 2 important classes is finished by poppe. Aggarwal and Ryoo conjointly categorised activity recognition into two sub-categories, the ‘single layer’ approach and ‘hierarchial’ approach. All these can be summarized as

|  |  |
| --- | --- |
| **Authors and year** | **Area of interest** |
| Aggarwal and cai(1999) | Human motion analysis and tracking from sinle and multi view data |
| Jaimes and sebe (2008) | Multimodal affective interaction analysis |
| Turaga et al (2008) | Categorization of actions and activities |
| Poppe(2010) | Action classification |
| Aggarwal and Ryoo (2011) | Gestures, human activities, actions and interactions analysis |
| Chen et al. (2013) | Human body part motion analysis |
| Ye et al. (2013) | Human activity analysis from skeletal poses |
| Aggarwal and xia (2014) | Human activity analysis from stereo, motion capture |
| Guo and lai (2014) | Understanding human activities from still images |

Later also there are many research works related to this that can be cited in papers like

Human action recognition with video data: Research and evaluation challenges [1]

In this paper, we got an overview of existing methods, their ability to recognize abnormal actions. This helped us to identify suitable methods available to address each challenge and it’s limitation.

3D convolution neural networks for human action recognition [2]

In this paper they developed a 3D CNN model for action recognition. In this motion information is captured from multiple adjacent frames.

Public acceptance of drones: Knowledge, attitudes and practice [3]

This article stated about drones and their applications, uses in safety of the public and surveillance.

Human physical activity recognition based on computer vision with deep learning model [4]

In this paper different techniques that can improve performance of neural networks and other activation functions are tested and many different activities are considered.

End to end learning of deep CNN for 3D human action recognition [5]

In this paper action recognition is done using deep convolution neural networks by transforming skeleton sequence into image.

Deep CNN for human action recognition using depth maps and postures [6]

We referred this mode to increase feature extraction for accurate action classification. In this paper 3 CNN channels are trained with different inputs to achieve this.

Convolution 2 stream network fusion for video action recognitions [7]

This model is used for incorporating motion information. It proposed many solutions for this. In this for video snippets they proposed a new convNet architecture for spatio temporal fusion.

Ensemble deep learning for skeleton based action recognition using temporal sliding LSTM network [8]

This is used for feature representation of skeleton joints to recognize human activities. In this paper relation between related activities and features is analysed by visualizing softmax features.

Human action recognition using transfer learning with deep representations [9]

Based on pretrained deep CNN model this paper gives a method for human action recognition using SVM and KNN classifiers and it is also confirmed by comparative analysis that proposed method gives superior performance in accuracy.

Drone-Action: An Outdoor Recorded Drone Video Dataset for Action Recognition [10]

This is one of the available related paper on our project that deals with action recognition from drone videos recorded at outdoors.

We collected the required information from all this and in addition we also worked on hyper parameters, estimators of random forest.

**4. Proposed Work**

This work embody knowing concerning machine learning, random forest model, boosting technique, hyper parameters, tuning them and every one. Since machine learning provides system the flexibility to find out things mechanically, during this it makes system to mechanically classify human activities. Random forest is one among the metric capacity unit method used for classification of tasks with textile and boosting techniques. In this we have an impulse to use boosting in order that machines square measure trained consecutive. In boosting one learns from different that successively boosts the training. we used this by putting in xgboost package. The projected ways also are enforced to avoid overfitting in random forest.

**4.1 Dataset Collection**

The dataset required by our project is created by us only. Initially we recorded videos of different activities performed by 9 members including both abnormal and normal activities. The 14 activities performed are kicking, throwing, swinging pipe, strangling, sitting, punching, sos, falling, walking, standing, waving, pushing, running, jumping. We made 270 video sequences of all these activities and divided each second of video sequence into 60 frames. The sequencing of video is done by using vlc media player and we had written code for pre-processing work and csv file extraction in module demo\_video. These frames from the pre-processed video sequences are datasets for pose estimation algorithms. Later we make csv file with coordinate points for key point abstraction of pose estimation. This csv file act as dataset for our random forest model. It contains coordinate points from frames and also each one is labelled accordingly. There will be some key points which will not be captured by drone at some angles. During the formation of CSV file we made those hidden points as -1 and for better accuracy we replaced those -1’s with previous value while pre-processing.

Refer Fig:2 for sample table of csv file

**4.2 Pose estimation**

The system takes a color image of size as input and produces, as output, the 2D locations of anatomical key points for every person within the image. First, a feedforward network simultaneously predicts a set of 2D conﬁdence maps of body part locations and a set of 2D vector ﬁelds of part afﬁnities, which encode the degree of association between parts.  The network is split into two branches: the top branch, shown in beige, predicts the conﬁdence maps, and therefore the bottom branch, shown in blue, predicts the afﬁnity ﬁelds. The image is ﬁrst analyzed by a convolutional network, generating a group of feature maps that’s input to the ﬁrst stage of every branch.

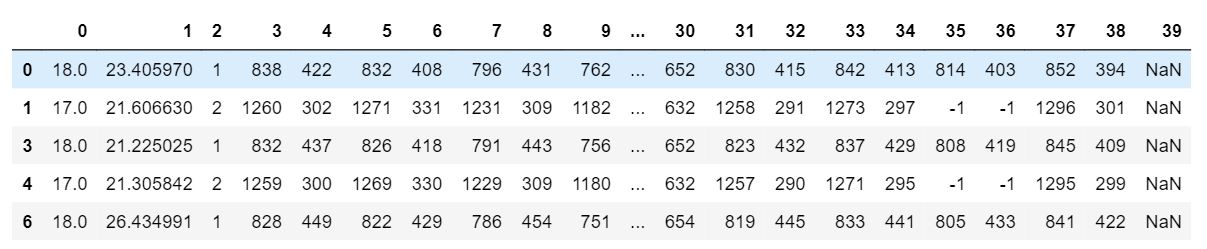


Fig: real time image for kicking

 Fig: key point abstraction of real time image

At the ﬁrst stage, the network produces a set of detection conﬁdence maps and a set of part afﬁnity ﬁelds. To evaluate loss function during training, we generate the ground truth conﬁdence maps from the annotated 2D key points. Each conﬁdence map may be a 2D representation of the assumption that a body part occurs at each pixel location. Ideally, if one person occurs within the image, one peak should exist in each conﬁdence map if the corresponding part is visible; if multiple people occur, there should be a peak like each visible part for every person. We generate individual conﬁdence maps for every person. We take the utmost of the conﬁdence maps rather than the typical in order that the precision of accessible peaks remains distinct. The set A has conﬁdence maps, one per part.

A 2D vector encodes the direction that points from one part of the limb to the opposite. Each sort of limb features a corresponding afﬁnity ﬁeld joining its two associated body parts. The set B has vector ﬁelds, one per limb1, each image location within the set encodes a 2D vector. Finally, the conﬁdence maps and therefore the afﬁnity ﬁelds are parsed by greedy inference to output the 2D key points for all people within the image. The part afﬁnity may be a 2D vector ﬁeld for every limb, it preserves both location and orientation information across the region of support of the limb.



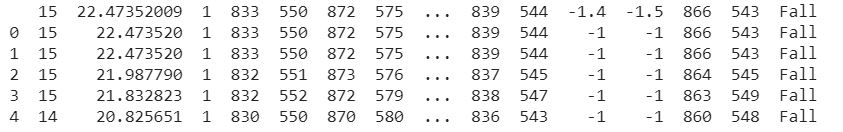


Fig:2 sample table of CSV file

**4.3 Action classification**

Human action recognition, classiﬁcation and understanding in videos have been a signiﬁcant research domain in computer vision. In order to attain a physical action, the person should interact and give feedback to the environment using his/her head, hands, arms, legs, bodies. Different joints position express different actions. Hence, several joints movement of human body can be considered as human activity. video-based human actions consist of a sequence of inert postures. There are two different approaches to analyse the human pose such as, top-down approaches and bottom-up approaches. Each method has its own characteristics. In our research, we prefer to utilize real-time human pose estimation method.  We use the 2D coordinates obtained by using real-time human pose estimation approach as the cues to action classiﬁcation. In the proposed approach, the 2D coordinates of a person in each frame in the video clips are extracted and accumulated as matrix. Then these accumulated coordinates are used to feed the Random forest classification model to classify the human activities. There are totally 14 classes to predict -punch, kick, strangling, throwing, swinging pipe, sitting, sos, falling, walking, standing, waving, pushing, running, jumping. It creates decision trees and compare the votes collected from each tree to predict the action. There are two well-known methods in Random forest which is boosting and bagging. In boosting successive trees give extra weight to points incorrectly predicted by earlier predictors. within the end, a weighted vote is taken for prediction. In bagging, successive trees don’t depend upon earlier trees. They independently constructed employing a bootstrap sample of the info set. Within the end, an easy majority vote is taken for prediction. First, bootstrap samples are created from the original data. For each of bootstrap samples, an unpruned classiﬁcation or regression tree is grown, with the following modiﬁcation: at every node, instead of choosing the best split among all predictors, randomly sample of the predictors and choose the best split from among those variables. Predict new data by aggregating the predictions of the trees (i.e., majority votes).

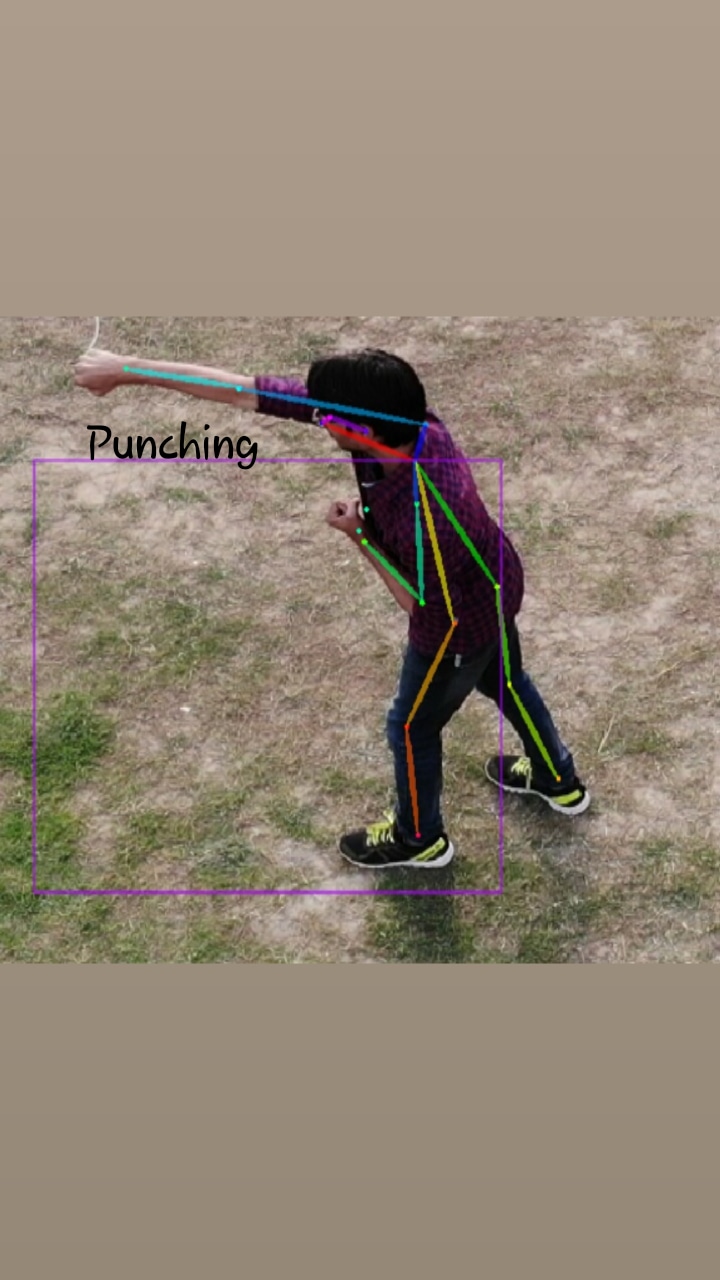
****

Fig: action recognition after key point abstraction

**5. Experiment and results**

**5.1 Data set**

We made 270 video sequences of all these activities and divided each second of video sequence into 60 frames as data. The data is divided in to training data and testing data set in the ratio of 8:2. These frames from the pre-processed video sequences are datasets for pose estimation algorithms. A CSV file was extracted from every video containing the key points for each frame of the corresponding video. For more accuracy all the hidden points were replaced by their corresponding previous values which resulted in better accuracy.

**5.2 Feature extraction**

The features are extracted by using the key points. This algorithm detects 18 key points totally from body parts.The co-ordinates of the key points are extracted into the csv file. These features were further used in the random forest model for forming the decision tree.

**5.3 classification results**

The coordinates of the key points are given as input to the Random forest classification algorithm There are totally 14 classes to predict and our model predicts with the training accuracy of 99% and testing accuracy of 96.6%

**5.3.1 Performance Matric**

**Accuracy:** Measure of total number of predictions that are perfectly classified.

**Precision:** Measure of specific cases expected based on confident.

**Recall:** It is similar to true positive rate.

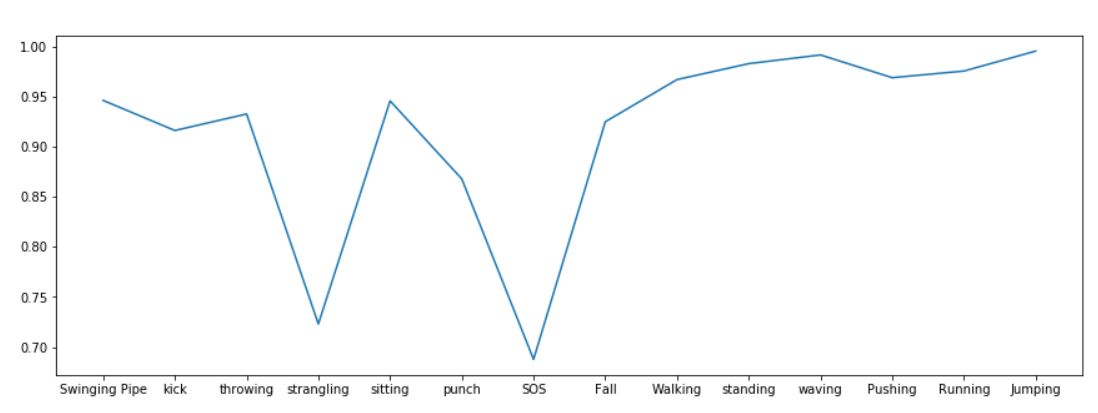
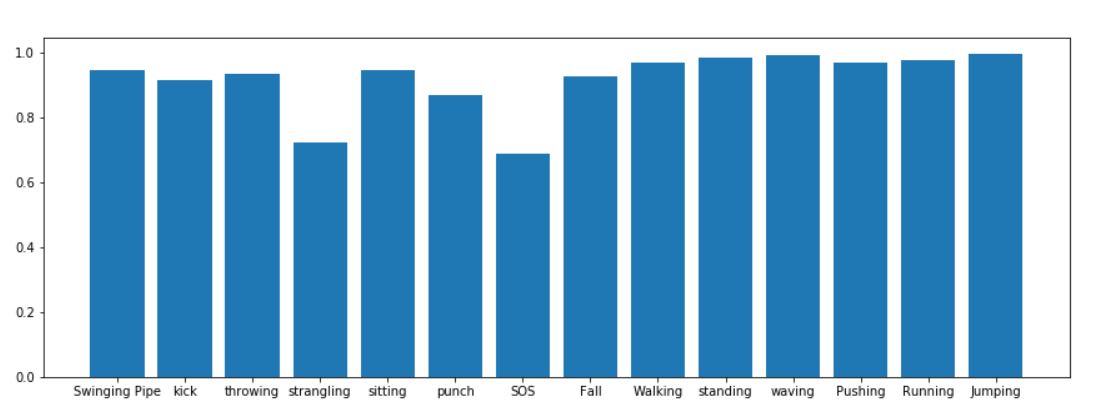
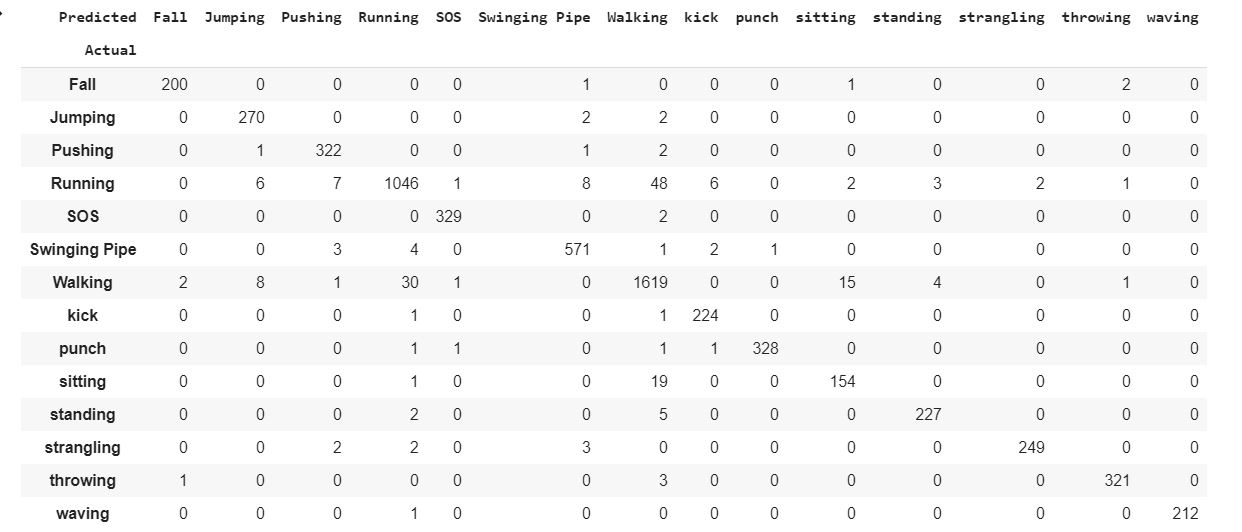
 

Fig: line graph depicting accuracies of each activity Fig: Bar graph depicting the accuracies of each activity

**5.3.2 Analysis**

It consists of data about known class and certain class. Here rows of matrix style the known defined values and posts of the matrix describe the expected values. The traverse values are classified are classified seamlessly and therefore the off-diagonal values are falsely classified. The Resulted confusion matrix for our activities i



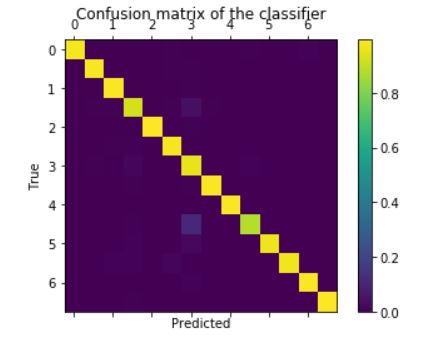


Fig: confusion matrix considering 7 activities

**6. Conclusion**

Action recognition has received a lot of interest because of the various research challenges that haven’t been satisfactorily self-addressed. Another vital reason is that the large vary of potential application like surveillance and human computer interaction etc. In this we have presented an efficient Human action recognizer which mainly classifies between violent and normal day to day activities. Pose estimation model and Action classification were used. The key points are identified using the pose estimation model and the features are extracted to csv file. The extracted features in the csv file are used to train the Random Forest classification model. The action classiﬁcation results of our proposed approach using six different types of human actions based on our dataset produce very good performance, which is higher than the other competitive approaches. In future, we would like to add more features to recognize with better accuracy. It can be also used for drone surveillance and crowd violence detection.

**7.References**

[1] Ramanathan, Manoj, Wei-Yun Yau, and Eam Khwang Teoh. "Human action recognition with video data: research and evaluation challenges." *IEEE Transactions on Human-Machine Systems* 44.5 (2014): 650-663.

[2] Ji, Shuiwang, et al. "3D convolutional neural networks for human action recognition." *IEEE transactions on pattern analysis and machine intelligence* 35.1 (2012): 221-231.

[3] Aydin, Burchan. "Public acceptance of drones: Knowledge, attitudes, and practice." *Technology in Society* 59 (2019): 101180.

[4] Mo, Lingfei, et al. "Human physical activity recognition based on computer vision with deep learning model." *2016 IEEE International Instrumentation and Measurement Technology Conference Proceedings*. IEEE, 2016.

[5] Li, Chao, et al. "End-to-end learning of deep convolutional neural network for 3D human action recognition." *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2017.

[6] Megavannan, Vennila, Bhuvnesh Agarwal, and R. Venkatesh Babu. "Human action recognition using depth maps." *2012 International Conference on Signal Processing and Communications (SPCOM)*. IEEE, 2012.

[7] Feichtenhofer, Christoph, Axel Pinz, and Andrew Zisserman. "Convolutional two-stream network fusion for video action recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[8] Lee, Inwoong, et al. "Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks." *Proceedings of the IEEE International Conference on Computer Vision*. 2017.

[9] Sargano, Allah Bux, et al. "Human action recognition using transfer learning with deep representations." *2017 International joint conference on neural networks (IJCNN)*. IEEE, 2017.

|  |
| --- |
|  |
| Bux, Allah, et al. "Human action recognition using transfer learning with deep representations." (2017).  [10] Perera, Asanka G., Yee Wei Law, and Javaan Chahl. "  Drone Action: An Outdoor Recorded Drone Video Dataset for ActionRecognition." *Drones* 3.4 (2019): 82. |

|  |  |
| --- | --- |
|  |  |

**8. Video Presentation**

<https://www.youtube.com/watch?v=fqXsf36NzQg>