ISSN: 0976-9102 (ONLINE)
DOI: 10.21917/ijivp.2022.0382

A ROBUST METHOD FOR HUMAN ACTION RECOGNITION IN VIDEO STREAMS USING SKELETON GRAPH BASED CNN

K.L. Bhagya Jyothi[1] and Vasudeva[2]

[1]Department of Computer Science and Engineering, KVG College of Engineering, India
[2]Department of Information Science, NITTE University, India

Abstract

Understanding the action of human plays an important role in public gatherings and recognition of human action is a major problem which leads to analysis the human activities. In many years, people are interested for detecting the human activity. Human behavior analysis are used in many areas like video surveillance, banks to increase public security. To detect the human behavior from the videos, essential features are to be detected. The major challenge in human action recognition is to generate the required features significance changes occurred in human action. Nowadays skeleton data-based action detection becoming more popular. In order to counterpart such limitations, this paper brings a method using Skeleton Graph based deep learning convolutional neural network. The proposed method gives accuracy of 0.93.

Keywords:

Human Action Recognition, Skeletization, Convolutional Neural Network, Skeleton Graph

1. INTRODUCTION

Recent year's human action recognition frequently used subject in the area of computer vision. Detecting the human behavioural movements from the video frames provides the method for classifying the one particular event such as walking, running, jumping, interacting with computer, and person-person interaction. Some activities like riding bike, playing some musical instruments, and group interactions may require some additional complex methods to detect movements in video [1].

Public safety is becoming more important nowadays for people from all aspects of life. Detecting human movements also plays important role in detecting some abnormalities in videos. With the new innovative methods in the video monitoring systems the detection the possible risks in the human actions are becoming important for public safety. The recognition of those activities with some methods based on video may not be accurate because of some number of noises. Hence because of these limitations encourages the researchers to handle the difficulties of human behavioral analysis in videos. The human pose estimation is a process which deals with detecting the parameters of a pose of a human body [2]. There are lots of efforts have been made in each method independently in different aspects and their effects on action recognition is still unknown.

The human behaviour detection of actions is a major problem because it may contain some disturbances, noises and filled with some background. Sometimes slight changes in actions length, the variance in the interclass, and various similarities in the class [3] which increases difficulty in recognizing the action. To cope up these difficulties the researchers inventing new types of method. The RGB images are normally used for human action detection. Normally there are some features like histogram-oriented features, optical flows obtained from the color images to detect the human movements. Most of the techniques for the human action recognition, which is based on content data retrieval, machine-human interaction, intelligent video surveillance, and gaming systems [5]. There are many complex issues to be solved. Like some sport actions like boxing, kick boxing and, wrestling getting more attentions nowadays, because of lack of datasets

Human action recognition in video streams is highly complicated task in computer vision. This research is to recognize the activities which involves multiple humans and gives the needful information to help in IoT applications [6]. In video surveillance because of sensor cameras and cloud data makes the s data easily taken and shared with other systems. In video surveillance identifying the action in all the video frames in the sequences are important. Some of the frames contain less important information which may give some misleading information with other actions, where some frames may give more correct information [7].

In most of the existing recognition algorithms usually does not take the depth information as well as key actions frames in recognition gives low efficiency and accuracy in the results [8]. But in recent years a number of 3D technology developed and quality of traditional based techniques are improved a lot. Traditional Hand-crafted feature are usually for simple techniques which cannot handle complex scenarios. Also, traditional methods cannot detect the actions which is of real time videos which has some complex backgrounds.

Human posture detection is a frequently used topic in current literature in various applications. There are a variety of algorithms and techniques that gives the solution to this task with high accuracy.

Today deep learning methods used mainly in neural networks which gives contains several layers for the classification. Usually, the techniques related to deep learning are more effective than traditional methods. Neural network-based techniques provide more accurate outcome in human behavior analysis systems. It has more ability to obtain the features. The deep neural methods are used to handle the complex human actions which is more efficient than the traditional methods. Deep learning-based CNN are widely used in different computer vision related techniques [3].

Deep learning-based CNN plays a major part in obtaining features from original videos. In convolutional neural networks compared to other traditional systems there will not be any pre-processing steps before giving the image to network. In CNN method, the extracted frames are given to CNN then features are extracted from one or more layers. In recent years there are many advancements in CNN model makes the accuracy when it is compared to other traditional methods. Currently the deep learning methods gives the superior performance where handcrafted features are used merely. CNN based deep techniques nowadays considers only spatial data and avoids the temporal data which plays an important role in human recognition system.

Convolution neural networks plays important role in recent years. Because of excellent performance in image recognition and detection, CNN are used extensively used in image detection and classification. Because of huge success in image recognition, CNN are used in many other fields of artificial intelligence

Nowadays there are many works which are based on skeletal data of image. In this work skeleton data is directly given to the neural network model and model will extract the features from the data. Each skeleton image contains joints which is shown in the Fig.1 [21]. Human pose can express as either 2D/3D skeleton representation which makes the algorithm for better classification of human action.

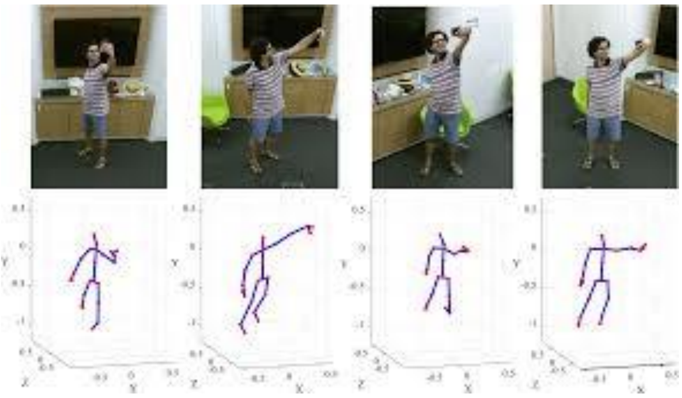


Fig.1. Human actions as skeleton mapping

In skeleton-based techniques features are encoded as angle information of joints and displacement vector of joints. In skeleton-based image representation the storage skeleton is very less compared to RGB images. Here only joint information's are stored. Normally total number of joints will be not more than 30. In RGB representation it requires space for all the pixels. Hence storage cost is less in skeleton-based approach.

Normally in Conventional methods the skeleton-based technique mainly focuses on extracting the handcrafted features which are used to represent the skeleton image. But currently because of neural networks, extracting features become effective compared to traditional systems. Compared to traditional systems deep learning methods methods are more efficient.

When we compared with different convolutional techniques, the skeleton-based action detection with CNN is getting more and more famous because of robust in complicated

The major motive of this work is to build a robust method for action detection of human using the CNN with Skeleton Graphs Technique. In this method video is taken as input which is then generated as frames. Once the frames are generated Skeleton sequences are generated in which features of the skeleton are generated efficiently. Overview of the proposed method is shown in the below Fig.2.

Our proposed approach has two steps. Human skeletal construction and CNN model for human action recognition. Normally human gestures consist data information in terms of spatial joints and temporal frames and 3D information when it contains depth information.

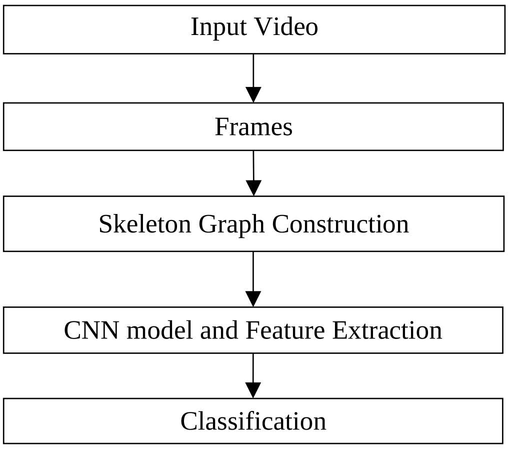


Fig.2. Outline of the Proposed Method

In this human skeleton image extracted using some skeletisation techniques. In this work human skeleton is modelled as set of joints. These joints are represented as Cartesian coordinates. The joints of human are measured with respect to measure of neck and the torso joint distances.

In this technique video is converted as frames. From each frame Skelton image is taken and considered as independent graph. Total number of skeleton join will be size of graph and physical connections will be the graph links. Each graph is expressed as $G = (V,E)$, here V represents node set which consists the joints of the body, which is of either 3D or 2D coordinates .Here the E is the set of edges in which it consists links of joints.. Normally, edges will be represented as an adjacency matrix A. Here $A_{ij} = 1$ which denotes that the link exists in joint i and j, and $A_{ij} = 0$ is zero for null edges. Once the skeletal graph is constructed then it is fed to the CNN for the action recognition.

1.1 MOTIVATION

In this section we have presented a few works which are based skeletal image data which are used in deep learning techniques for human action recognition. Nowadays researchers are showing more interest on graph data in image processing. They are used in many applications like chemical factories, e-commerce, networks applications etc.

2. LITERATURE SURVEY

There are many existing methods are as follows: in [8] this model which has three steps image silhouette, 3D Model, Skeletal Graph construction, labelled graph, Skeleton structure. In this paper, we have given an idea about generation of human pose as skeletal graph. This graph is represented as a tree like structure which is like human body. This structure can be used by the algorithm to detect the features. A given a method based on directed graph neural network which extracts joints information, bones and their relationships and makes the human action prediction [4].

In [10] silhouette images are used as 2D image which is used to create a skeletal graph and gives the labels for the end nodes using graph matching technique. In Another paper [11] uses a technique called Shock

Graph which is a representation of the skeleton shape for Directed Acyclic Graph. Here skeletal graph is represented based on the local difference of the radius function. Here data base for image is constructed by appropriate image acquiring technique, later these images are transformed into binary images. Later Skeletonization algorithm used for labeling process.

Table.1. Different Skeleton based Approaches

Sl. No	Dataset Used
1	HiEve
2	PoseC3D
3	UOW Online Action 3D
4	NTU-60 and NTU-120
5	MSR Action3D, MSR DailyActivity3D, Huawei/3DLife-2013
6	NTU RGB+D, UT-Kinect and SYSU 3D
7	NTU RGB+D and UWA3DII.
8	CMU motion capture
9	NTU RGB+D, NTU RGB+D 120, and NW-UCLA

In [12] process starts from generating skeleton joints and then calculates feature vector for each action. In the next step multiclass machine learning algorithm are used, in this process different activity is represented as different class, later it will be used for classification purpose. This technique has manly four steps which are as follows

- **Posture Features Extraction**: The skeleton joints coordinates are considered to measure the performance of the feature vectors which expresses the postures of the human.
- **Postures Selection**: From each activity essential postures are considered.
- **Activity Features Computation**: In the next step a feature vector which is considered for the entire activity is generated and applied for classification.
- **Classification**: A multiclass SVM implemented is used for classification.

The authors in [20] developed a model which is based on hierarchical neural network which operates with skeleton-based approach. There are many research works were published based on skeleton graph method in human action recognition which is shown Table.1.

2.1 MAJOR CHALLENGES

Most of the research work are still lacking the following factors in Human Action Recognition: 1) Multimodal visual Perception, 2) Variations in actions Fast Action Detection, 3) More concentrate on Real Scenarios, 4) Rapid Movements in the video, 5) Noise in the video, 6) Movements in background and 7) Multiple people movements.

3. PROPOSED HUMAN ACTION RECOGNITION USING SKELETON GRAPH WITH CONVOLUTIONAL NETWORK

In skeletonization process the input image is converted and represented as stick like structure. Skeleton image is viewed as 2D structure which preserves all the details of the silhouette. The main aim is to develop effective model for human action detection-based skeleton graph with CNN model. In our work the system consists the following steps which are as follows, frame generation, converting the frame image into 2D/3D structure, then skeletons of image are obtained using some skeletonization algorithm. Then skeleton is labelled and given as input to the trained CNN model for action recognition.

3.1 ACQUIRING THE INPUT VIDEO

Consider *D* be the dataset which contain n number of samples. These samples are expressed as,

$$D=\{V_1,V_2,..V_i,....V_n\} \tag{1}$$

3.2 EXTRACTION OF FRAMES

Once the video samples are acquired, each sample is converted as number of frames for the next processing. Each video *V*i is given to next step to extract frames. Once the frames are extracted it is expressed as set *E*,

$$E=\{E_1,E_2,....E_j,...E_m\};E_j\in V_i \tag{2}$$

In above equation E_j represents video frame V_i and *m* denotes total count of generated frames in the video such that *j* = {1, 2, 3….*m*}.

3.3 SKELETON GRAPH CONSTRUCTION AND FEATURE EXTRACTION

The generated video frames *E* is given as input to skeletisation algorithm for extracting the features. In the convolutional networks based on graphs, skeleton of the image data is expressed as a structure *G* (*V*, *E*), vertices will be the body joints which is denoted by the symbol *V*. Bones are represented as edges which is denoted by *E*. If two frames contain same joints then it is connected by edges. Skeleton joints are used to represent the structure which is used to extract necessary information for behaviour classification.

3.4 HUMAN ACTION RECOGNITION USING SKELETON GRAPH BASED CNN MODEL

The generated feature *F* is fed as input to Convolutional Neural Network for human action recognition to identify actions of humans from video frames. The CNN is a kind of network which contains multiple layers. The extracted feature F_j is given as input to the CNN, which consists a state vector input vector of a hidden layers and it is expressed in the following Fig.3:

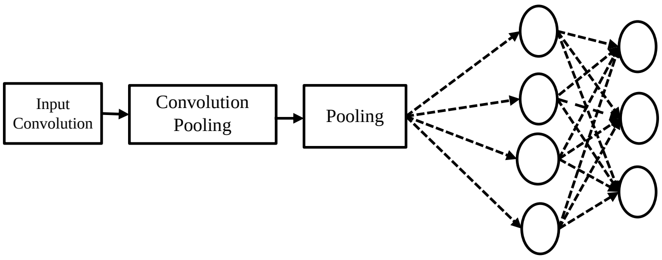


Fig.3. Overview of the CNN Model

The convolution layer obtains the generated features from the input images. The operation of convolution layer is conducted between the image and a filter of a particular size *M*×*M*. Dot product is performed for the filter and some portions of the input image to the size of the filter. Next convolutional layer output is given to the pooling layer which decrease the size of the convolved feature map to decrease the computational costs. The next layer is fully connected layer which consists the weights and biases also the neurons which is used to connect the different neurons of the different layers. This layer is placed just before the output layer.

Feature vector are generated in each layer is connected to the every generated in input layer. The total number of feature vectors are increased as the number of layers is increased. In 2D convolution features are extracted from the local neighbourhood feature vector from the previous layer. Here additive bias function is applied then the result is given to the sigmoid function. In our model CNN uses the skeleton data which is more efficient compared to other methods. In our methods two convolutional layers and two Max Pooling layers are used. Activation function we have used in the layer is PReLU (parametric Rectified Linear Unit). In this activation function only, positive values are given and negative values are converted as zero. The PReLU can be expressed as:

$$F\left(y\right)=\begin{cases} y_i & \text{if } y_i>0 \\ a_i y_i & \text{if } y_i\le 0 \end{cases}$$

Here a_i is used to control the deviation of negative part. The structure of PReLU function is shown below in the Fig.4.

Fig.4. Activation function top (ReLU) and down (PReLU)

4. RESULTS AND DISCUSSION

This section describes the results of developed algorithm for human action recognition

4.1 EXPERIMENTAL SETUP

The proposed model is implemented in MATLAB tool with KTH action dataset.

4.2 DATASET DESCRIPTION

To experiment this work KTH Action dataset is used.

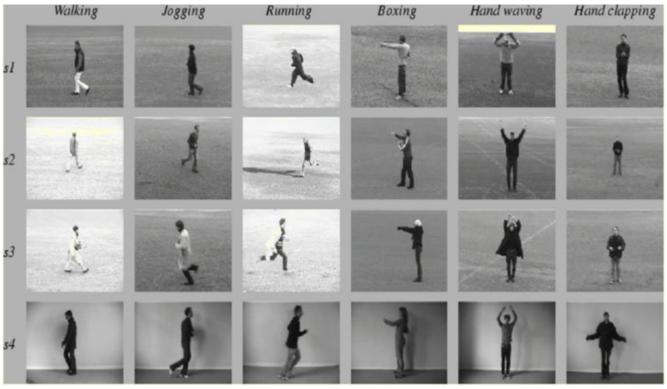


Fig.5. Sample actions in KTH Data set

The KTH action dataset is considered as frequently used datasets, KTH data set consists mainly 6 varieties of actions namely jogging, walking, running hand clapping, hand-waving [13]. To measure the performance every action is done by 25 different type's persons. The Fig.5 shows the various action categories in KTH Data set.

4.3 EVALUATION METRICS

The algorithm is studied by applying the performance measurements like accuracy, sensitivity and specificity. The confusion matrix is shown in Table.2.

Table.2. Confusion matrix (%) for accuracy for the KTH action dataset

Class	Boxing	Hand clapping	Hand Waving	Jogging	Running	Walking
Boxing	0.93	0.13	0.02	0.00	0.04	0.00
Hand clapping	0.00	0.92	0.08	0.01	0.00	0.00
Hand Waving	0.00	0.10	0.94	0.00	0.00	0.00
Jogging	0.03	0.00	0.04	0.93	0.05	0.08
Running	0.02	0.03	0.10	0.05	0.92	0.02
Walking	0.00	0.00	0.10	0.00	0.00	0.91

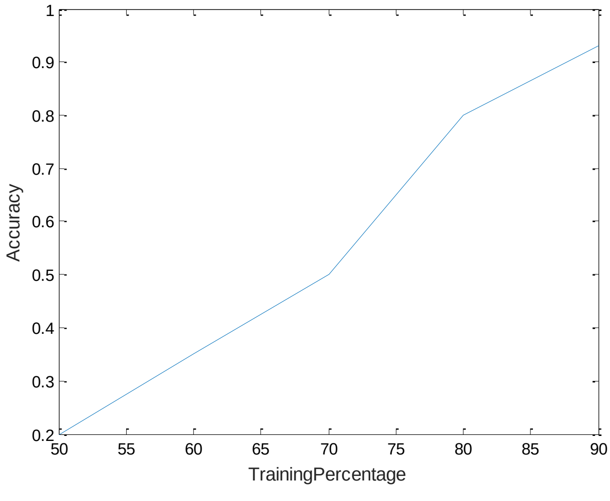


Fig.5(a). Analysis based on Accuracy

In the above confusion matrix, we can see various values are noted for various types of actions. The proposed work is tested and accuracy measured is 93% which is quite reasonable compared to other methods. The graphs are plotted with respect to accuracy, sensitivity and specificity. The training samples were taken 50%, 60%, 70%, 80%, 90%. The number of iterations were taken as the analysis of these measurements are shown in the Fig.5(a)-Fig.5(c), respectively.

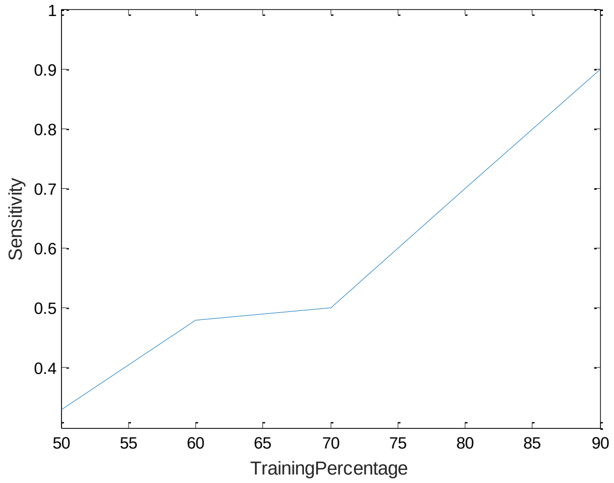


Fig.5(b). Analysis based on Sensitivity

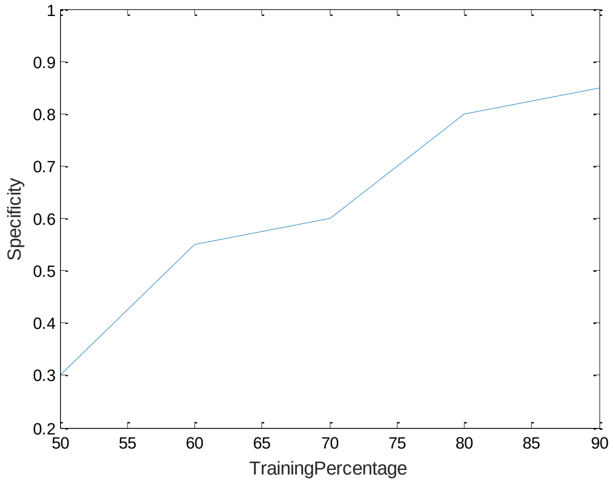
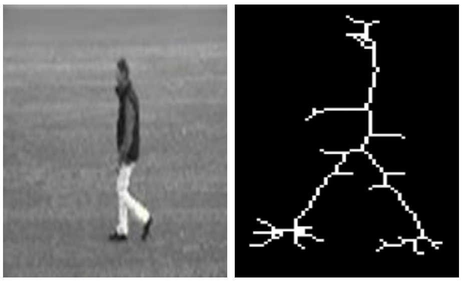
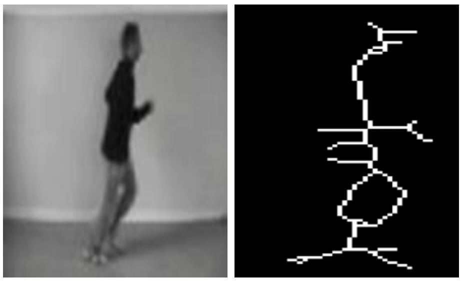


Fig.5(c). Analysis based on Specificity



(a) (b)



(c) (d)

Fig.6. Experimental results, a) Input video for walking b) Skeleton graph constructed for image frame for walking c) Input video for running d) Skeleton graph constructed for image frame for running

4.4 COMPARATIVE DISCUSSION

The Table.3 gives the summary of some of the previous works compared with the proposed work. Our proposed method is reasonably better compared with the other methods which is discussed in the table. Hence, we can say that the proposed method efficient compared to other methods. Our proposed method achieved the accuracy 93% with the KTH action data set.

Table.3. Comparative discussion

Authors	Accuracy (%)
Nowozin et al. [14]	87.04
Neibles et al. [15]	81.50
Krizhevsky [16]	87.49
Arunnehrua et al. [17]	94.9
Antonik, P et al. [18]	91.3
Zahraa Salim David et al. [19]	90
Proposed Skeleton Graph based CNN	93

5. CONCLUSION

In videos human behavior analysis and detection becomes popular topic in nowadays in video surveillance systems, traffic monitoring systems, automatic vehicle system, pedestrian monitoring system and robotics. In human action detection in video each frame plays different roles in feature learning. There are several techniques are developed in this area. Recently deep learning-based techniques becomes more and more popular compared other traditional methods. In deep learning CNNs are widely used to detection of human action. But human behavior analysis systems face different challenges because of various issues like multiple person interaction, noises in videos, multiple actions etc. To address these issues, we proposed an effective mechanism for human action detection using Skeleton graph with Deep Convolution Neural Network. Our proposed approach contains frame generation, Skeleton graph construction and feature learning using CNN. The proposed achieved a maximum accuracy of 0.93.

REFERENCES

[1] Chengfei Wu and Zixuan Cheng, "A Novel Detection Framework for Detecting Abnormal", *Mathematical Problems in Engineering*, Vol. 2020, pp. 1-9, 2020.

[2] Matthias Straka., Stefan Hauswiesner, Matthias Ruther and Horst Bischof, "Skeletal Graph Based Human Pose Estimation in Real-Time", *Proceedings of British Conference on Machine Vision*, pp. 1-6, 2011.

[3] Lei Shi, Yifan Zhang, Jian Cheng and Hanqing Lu, "Skeleton-Based Action Recognition with Directed Graph Neural Networks", *Proceedings of International Conference on Computer Vision*, pp. 1-5, 2018.

[4] M. Vrigkas and A.A. Ioannis, "A Review of Human Activity Recognition Methods", *Sensors*, Vol. 19, No. 17, pp. 3680-3688, 2019.

[5] Ali Mottaghi, Mohsen Soryani and Hamid Seifi, "Action Recognition in Freestyle Wrestling using Silhouette-Skeleton Features", *Engineering Science and Technology*, Vol. 23, pp. 921-930, 2020.

[6] M. Feng and J. Meunier, "Skeleton Graph-Neural-Network-Based Human Action Recognition: A Survey", *Sensors*, Vol. 22, pp. 2091-2099, 2022.

[7] K. Zhou, "Skeleton Based Abnormal Behavior Recognition using Spatio-Temporal Convolution and Attention-Based LSTM", *Procedia Computer Science*, Vol. 174, pp. 424-432, 2021.

[8] X. Zhen and L. Shao, "Action Recognition Via Spatio-Temporal Local Features: A Comprehensive Study", *Image and Vision Computing*, Vol. 50, pp. 1-13, 2016.

[9] M. Straka, Stefan Hauswiesner and Matthias Ruther, "Skeletal Graph based Human Pose Estimation", *Proceedings of British Conference on Machine Vision*, pp. 1-12, 2011.

[10] Nicolas Thome, Djamel Merad and Serge Miguet, "Human Body Part Labeling and Tracking using Graph Matching Theory", *Proceedings of IEEE Conference on Video and Signal Based Surveillance*, pp. 1-14, 2006.

[11] P.M. Pandit and S.G. Akojwar, "Skeletonization and Classification by Bayesian Classifier Algorithm for Object Recognition", *International Refereed Journal of Engineering and Science*, Vol. 2, No. 5, pp. 24-31, 2013.

[12] Meng Li and Q. Sun, "3DSkeletalHumanAction Recognition Using a CNN Fusion Model", *Mathematical Problems in Engineering*, Vol. 2021, pp. 1-9, 2021.

[13] S. Nowozin, G. Bakir and K. Tsuda, "Discriminative Subsequence Mining for Action Classification", *Proceedings of International Conference on Computer Vision*, pp. 1-12, 2007.

[14] J.C. Niebles, H. Wang and L. Fei-Fei, "Unsupervised Learning of Human Action Categories using Spatial-Temporal Words", *Proceedings of British Conference on Machine Vision*, pp. 1-7, 2006.

[15] A. Krizhevsky, I. Sutskever and G.E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks", *Proceedings of Annual Conference on Advances in Neural Information Processing Systems*, pp. 1-13, 2021.

[16] J. Arunnehru,, G. Chamundeeswari and S. Prasanna Bharathi, "Human Action Recognition using 3D Convolutional Neural Networks with 3D Motion Cuboids in Surveillance Videos", *Proceedings of International Conference on Robotics and Smart Manufacturing*, Vol. 133, pp. 471-477, 2018.

[17] P. Antonik, N. Marsal and D. Brunner, "Human Action Recognition with a Large-Scale Brain-Inspired Photonic Computer", *Nature Machine Intelligence*, Vol. 1, pp. 530-537, 2019.

[18] Zahraa Salim David and Amel Hussain Abbas, "Human Action Recogntion using Interest Point Detector with Kth Dataset", *International Journal of Civil Engineering and Technology*, Vol. 10, No. 4, pp. 333-343, 2019.

[19] Y. Du, W. Wang and L. Wang, "Hierarchical Recurrent Neural Network for Skeleton based Action Recognition", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 1110-1118, 2015.

[20] P. Zhang, "View Adaptive Neural Networks for High Performance Skeleton-Based Human Action Recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 8, pp. 1963-1978, 2019.

[21] Noor Almaadeed, "A Novel Approach for Robust Multi Human Action Detection and Recognition based on 3D-Dimentional Convolutional Neural Networks", *Proceedings of British Conference on Machine Vision*, pp. 1-12, 2019.