Comparative Analysis of AI-powered Approaches for Skeleton-based Child and Adult Action Recognition in Multi-person Environment

Conference Paper · March 2022

DOI: 10.1109/CSASE51777.2022.9759717

CITATIONS
2

READS
256

1 author:

Malithi Mithsara
Southern Illinois University Carbondale
4 PUBLICATIONS **7** CITATIONS

SEE PROFILE

Comparative Analysis of AI-powered Approaches for Skeleton-based Child and Adult Action Recognition in Multi-person Environment

Mithsara. W.K.M
Department of Computing
Rajarata University of Sri Lanka
Mihintale, Sri Lanka
malithim@as.rjt.ac.lk

2022 International Conference on Computer Science and Software Engineering (CSASE) | 978-1-6654-2632-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/CSASE51777.2022.9759717

***Abstract*—Machine learning on graphs is a unique framework structure to handle multiple objects. A lot of studies have gone into the topic of action recognition. Among them, most action recognition systems nowadays are based on skeletons. Now, GNN-based techniques are effective. In GNN-based techniques, the skeleton represents the graph, node as joints, and edges as the bones. As well as temporal relationship between the skeleton points could be obtained frame by frame. Most studies focus on detecting only one action for a single individual rather than several actions performed by multiple persons simultaneously in an untrimmed video in a well-segmented video. Many action recognition systems are focused on adult actions. However, no additional techniques for child action recognition. The action recognition of children is complex since there is no standard dataset for child action recognition. This study aims to determine a child's actions in a multi-person scenario. Initially, this method selects a minor domain action. Identify the child in the video using yolov5's custom object detection. Compare the ANN, 1DCNN, LSTM, and GNN (STGCN) with skeletal data to recognize the actions. The ANN method hasn't successfully determined the temporal relationship between frames. As a result, the GNN-based technique was employed in a multi-person situation to recognize behaviours. Adult skeleton data was taken from standards action datasets KTH and NTU-RGBD-120 using AlpaPose. Cropped YouTube videos are used for child action recognition. In terms of efficiency and accuracy, the GNN-based technique outperforms the others.**

Keywords—Child Action Recognition, Multi-person environment, GNN, Object Detection

I. INTRODUCTION

Machine learning on graphs is a one-of-a-kind framework with a structure to deal with multiple objects. A famous real-world illustration is a social network graph. Persons in a network are represented as nodes, and edges represent their connections. We want to encode pairwise properties between nodes, such as relationship strength or the number of shared friends, in this case of social network link prediction [1]. Action recognition is a well-known problem in computer vision. There are a variety of methods for recognizing individual actions and action localization. In a multi-person and multi-object situation, identifying actions might be challenging. Individuals should be conscious of whether they are involved in each action. Many issues in action identification, such as occlusion, lighting changes, location, recognizing human gestures in a single frame, and extracting frame-wise correlations, are difficult to overcome using typical methods [2]. Based on the types of input data, existing action recognition literature can be divided into two categories: image-based and skeleton-based approaches. The former includes the 3D skeleton from Microsoft Kinect [3] and the 2D skeletons from OpenPose and AlphaPose. The latter includes single-frame images [4], multi-frame video, and optical flow. The most important thing is the skeleton data is much smaller than the image data. The 3D skeletal data and image data in the NTURGB+D dataset, for example, are 5.8 and 136 GB, respectively. As can be seen, the skeletal data is 23 times smaller than the image data. As a result, skeleton-data training is faster than image-data training. [5] The study's primary goal is to identify the child in the video and detect the child's actions in a multi-person, multi-object environment using graph neural networks. The proposed approach first recognizes the child and adult in the video using a child-adult classification model and then identifies each child's behaviour in the multi-object environment.

II. RELATED STUDIES

A. Action Recognition

In computer vision, video action recognition is a novel framework. This field has a lot of research going on. Because of deep learning, we've seen significant progress in video action identification over the previous decade. However, we encountered unexpected difficulties, such as modelling long-range temporal information in videos, high processing costs, and incomparable outcomes due to dataset and evaluation protocol differences [6]. Earlier, Handcrafted features were used for understanding the video actions. But, handcrafted features have a high computational cost and are difficult to scale and implement [7]. As a result, researchers began to modify CNNs for video challenges as deep learning became more widespread [10]. Numerous studies focus on single action recognition in video, and only a few papers on multi-person action detection in a multi-object context. In recent decades, more studies have focused on skeleton-based activity recognition using deep learning techniques. Following is a discussion of the latter methods.

B. 3D ConvNet based

Tsai introduced the "Deep Learning-Based Real-Time Multiple-Person Action Recognition System" [8]. He used YOLOv3 to track down several persons who appeared on the scene. They used the Deep SORT algorithm to track the people and assigned an identity (I.D.) number. FaceNet is then utilized for face recognition to determine whether the I.D. exists to identify the person's name for display. Then,

using sliding window design and NMS, use the I3D architecture for real-time action recognition.

C. LSTM based

"RNN-Based Personalized Activity Recognition in Multi-person Environment Using RFID," proposed by Woo *et al*. [9]. Their paper presents personalized activity identification as a new research avenue. They suggest employing a recurrent neural network to develop a model for everyone. They also indicated that time-sliced and annotated data be collected seamlessly utilizing a graph-based event processing technique. Finally, they tested three R.N.N. designs with different unit types on actual data, including RNN, LSTM, and G.R.U. "Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition" was discovered by Ordóez and Roggen [10]. Using LSTM recurrent units and convolutional (that can naturally perform sensor fusion and does not require expert knowledge in feature building and openly depicts the temporal dynamics of feature activations). Two datasets were used to test their methods, one used in a public activity recognition competition. Their findings show that their framework beats rival deep non-recurrent networks-based action detection on the challenge dataset by an average of 4%, outperforming some previously reported results by up to 9%. In these two systems, sensor-based data were also employed for action recognition.

D. GNN based

The study "Spatial-Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition" was proposed by Yan *et al.* This method attempts to determine a person's action based on spatial and temporal factors. They represent the human skeleton as a graph, with joints as nodes and bones as edges and connections between each joint in each frame in a temporal manner. They estimate poses in videos and create spatial-temporal graphs from skeleton sequences. Multiple layers of spatial-temporal graph convolution (ST-GCN) will be applied, resulting in higher-level feature maps. The conventional Softmax classifier will then assign it to the appropriate action category [11].

The proposed ST-GCN outperforms the previous state-of-the-art skeleton-based model on two tough, large-scale datasets. (Kinetics, NTU-RGB+D) This method recognizes actions in a single person in the video frame. Those methods are used to recognize multi-person actions, but the STGCN method is only used to detect single-person actions. According to the survey, no technique employs GNN to recognize actions in a multi-person environment. As a result, the proposed approach used STGCN to recognize the child's behaviour when interacting with the parent in this study.

III. METHODOLOGY

As a first stage in the study, it's difficult to distinguish between the child and adult in the video. We must first identify the child's action in a multi-person scenario because we need to recognize the child's action. In this case, Yolov5 creates a model for separating children and adults. There are just a few strategies based on the classification of children and adults. Ince *et al*. presented the concept of "Child and Adult Classification Using Ratio of Head and Body Heights in Images." They were trying to identify pedestrians and classify them based on the head-to-body ratios of children and adults [12].

Fig. 1 shows the proposed architecture of this approach.

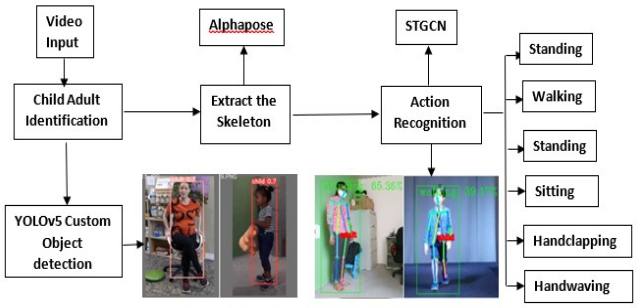

Fig 1. Proposed Architecture

Firstly, identify the child and adult given video input. Then extract the skeleton of the child and adult by using Alphapose. After each skeleton, go through the action recognition model to identify the actions of the child and adult.

A. Yolov5

Yolov5 is a deep learning-based approach to object detection. The former, such as RCNN, Fast RCNN, Faster RCNN, and mask RCNN, first find the locations of different objects in the image before recognizing them. The latter, such as YOLO and S.S.D., use a neural network to integrate both tasks. Object detection with excellent performance is achieved using YOLO (You Only Look Once) models. They can be used to detect objects in real-time using data streams. Yolov4 and Yolov5 used the deep SORT algorithm for detecting the objects. Deep SORT (simple online and real-time tracking) is a multiple real-time object tracking method proposed by Bewley.Yolov5 is more effective than Yolov1, Yolov2, Yolov3, and Yolov4. [13]

B. Object-Detection Model

Object detection is a method of identifying and locating objects in images and videos using computer vision.

Fig. 2. Annotation procedure

To identify the child in the video, I built a child-adult object detection model using 1200 photos of the infant and an Adult with 150 photographs for validation and testing. Finding a child and adult image dataset was tricky in this scenario. So, using YouTube videos connected to children and adults interacting, such as each video having children and adults and manually annotated as Fig. 2. The yolov5 was then used to train all the manually annotated pictures of children

and adults. After the training, the accuracy of the model increased to 85%. However, in this technique, the bounding boxes of both children and adults are primarily trained. Children's bounding boxes are smaller than adult bounding boxes. We can correctly detect the child and adult in the videos by resulting.

C. Action Recognition Model

In a multi-person environment, our system used the most effective method of STGCN [11] for child action recognition. Because this method used ANN, CNN, LSTM, and STGCN to recognize actions, the STGCN showed to be the most accurate. First, only focus on humans (children and adults) in the video without integrating other objects. Because this method began with a small set of actions such as standing, sitting, walking, jumping, handwaving, and handclapping before moving on to additional object-related activities. Skeletal data are taken from the AlphaPose used in all the approaches. Although AlphaPose is similar to OpenPose, it is more efficacious [14].

It was challenging to find single-action videos of children in the video, and there are no child action databases. As a result, this study gathered 20 videos of each child's movements from YouTube videos, which were chopped into action segments. (Standing, sitting, walking, clapping, waving hands, jumping). In addition, an adult action recognition model was developed, and then transfer- learning was used to create action recognition in children. The activities of standing and sitting are not included in the usual datasets. They were collected from YouTube. Furthermore, an adult action recognition model was created, and then transfer-learning was employed to help children gain action recognition. Standing and sitting activities are not included in the standard datasets. They were discovered on YouTube. AlpaPose was used to extract 26 skeleton points and a total of 7500 adult skeleton data for every single action (Standing, Sitting, Walking, Handclapping, Handwaving, Jumping) performed by the adults in the videos (50 videos per action), including x and y coordinates and confidence scores. A confidence score is a number that measures how certain each of the skeleton's joints is. The most common datasets, NTU-RGBD-120 and KTH, were employed in this study. A supervised learning technique was used to recognize actions.

Using AlphaPose, 1500 skeletal child data were extracted from the 20 videos, and two separate adult and child action recognition models were built. Here, transfer learning is used to assist the child in recognizing activities. Because of the low skeleton data from a child, an adult action recognition model was used to create a child action recognition model. The child, the action recognition model, delivers the same outcomes as the adult action recognition model (minor difference). Because this method only employed 2D data and did not consider the z coordinate. The model will be more accurate if we add depth information to it. However, in the future study, we want to integrate depth information to construct a more reliable action recognition model. After creating action recognition models for children and adults, the approach was tested separately using videos. Identify the child, feed the child skeleton into the child action recognition model, and then do the same with the adult skeleton.

C.1. ANN Approach

An artificial neural network (ANN) is a computing model of numerous processing modules that receive inputs and outputs [15]. The ANN was used to begin this study, and the accuracy of the ANN was compared. The input layer, two hidden layers, and the output layer make up the architecture of ANN-based action recognition. The 26 label skeleton data (only x and y coordinate of the skeleton joints) for each action is input into the input layer. The output layer outputs the six types of activities. However, there are no temporal connections between each video frame in this case.

C.2. 1DCNN Approach

When the temporal pattern between the frames is obtained, action recognition becomes more accurate. Because the video's actions are made up of frames. Therefore, one approach was identified as 1DCNN, and the design had temporal relationships between the blocks [16]. Some research has been conducted using 1DCNN to recognize actions. However, they all rely on sensor data [17] [18]. Rather than entering the frame by frame as in ANN, we can enter the collection of frames (one-person skeleton data) as a single block in 1DCNN. The max-pooling layer precedes the input layer and one 1DCNN hidden layer, flatten layer, and output layer in the architecture of 1DCNN-based action recognition. The skeletal data block is fed into the input layer, and the six actions are output into the output layer.

C.3. LSTM Approach

This approach is yet another method of determining the temporal link between frames. The research was proposed by Carrara *et al.* and is based on "LSTM-based real-time action recognition and prediction in human motion streams." They also sent ten frames per second to the LSTM layer [19]. Our LSTM Approach has an input layer (LSTM layer), two hidden layers (Fully connected layers), and the output layer. The input layer supplies the skeleton data (10 frames per second), and the six actions are output by the output layer, as in prior systems.

C.4. GNN Approach

In computer vision, the graph neural network is a novel framework. The skeleton can be determined as a graph. The nodes are the joints, and the edges are the bones. This approach utilized the STGCN [11] technique to detect the action in the video. The architecture is the same as the STGCN but includes a child-adult object detection model for testing child action in a multi-person environment.

D.Combining Object-Detection and Action Recognition

The skeleton data for children and adults are trained using ANN, 1DCNN, LSTM, and STGCN methods, with training divided into 0.8, validation 0.1, and testing 0.1. Following the training, this approach applied the trained action recognition model (which only recognizes a single person and action) to a video with many people, such as a toddler and an adult. We must feed each skeleton to our action recognition model one by one to use the action recognition model. So, first, we identify the child and adult using an object-detection model developed with yolov5 and their skeleton supplied into the model. Then we identify the child's and adult's behaviours.

E. AlpaPose

AlphaPose is similar to OpenPose, but according to some research, it is more effective. For the first time, OpenPose, a real-time multi-person human posture identification library, has detected the human body, foot, hand, and facial key points on single pictures. OpenPose can identify a total of 135 critical points [14]. The Alpha-Pose is based on top-down post-estimation. Top-down approaches are usually based on the precision of the person detector, according to the designers of this technology, because posture estimation is done in the area where the person is present. As a result of mistakes in localization and repetitive bounding box predictions, the pose extraction approach may perform sub-optimally. The developers designed a Symmetric Spatial Transformer Network (SSTN) Network to extract a high-quality human region from an incorrect bounding box to overcome this problem. A Single Person Posture Estimator (SPPE) was utilized to estimate the human stance skeleton for that individual in this extracted area. A Spatial De-Transformer Network (SDTN) was used to remap the human position back to the original picture coordinate system. The authors created a parametric pose Non-Maximum Suppression (NMS) strategy to cope with the problem of irrelevant pose deductions. A Pose Guided Proposals Generator has also been developed to help train the SPPE and SSTN networks by doubling training samples. Alpha-pose can be expanded to any combination of a person detection algorithm and an SPPE [14].

F. Datasets

NTU-RGBD-120: NTU-RGBD-120 [20] is a subset of NTU-RGBD, which is substantially larger and offers a broader range of ambient conditions, topics, camera angles, etc. It includes 114,480 video clips from 120 different actions. The subjects' ages range from 10 to 57, and their heights range from 1.3 to 1.9 meters. The dataset includes two criteria for evaluating action categorization performance: cross-subject and cross-setup. Fifty videos of adults handclapping and handwaving, and jumping actions were selected from this dataset to perform the action recognition.

KTH: The KTH Royal Institute of Technology started working on a non-trivial, publicly available dataset for action recognition in 2004. One of the most popular datasets is the KTH dataset, which comprises six actions: walk, jog, run, box, hand-wave, and hand clap. From this study, I chose 50 videos of people walking. [21]

Own Dataset for Child Actions: Finding child action datasets is tough. As a result, I used YouTube videos, cropped the activities, and created small videos with 30 frames per second. This dataset contains 20 videos for each action. (Handclapping, Handwaving, Sitting, Standing, Jumping, Walking) Here, the child's age is primarily considered between 1-and 6.

G. Skeleton dataset

The skeleton dataset that we utilized to train the action recognition model is an excel document with 42 columns containing video-name and frame-no and the x and y coordinates and the confidence score of the skeleton's 13 joint points, and the label of the action. Variables are video-name, frame-no, nose_x, nose_y, nose_s, LShoulder_x, LShoulder_y, Lshoulder_s, Rshoulder_x, Rshoulder_y, Rshoulder_s, LElbow_x, LElbow_y, Lelbow_s, Relbow_x, Relbow_y, Relbow_s, LWrist_x, LWrist_y, Lwrist_s, Rwrist_x, RWrist_y, Rwrist_s, LHip_x, LHip_y, Lhip_s, RHip_x, RHip_y, Rhip_s, LKnee_x, LKnee_y, Lknee_s, RKnee_x, RKnee_y, Rknee_s, LAnkle_x, LAnkle_y, Lankle_s, RAnkle_x, RAnkle_y, Rankle_s and the label of the action. AlpaPose was used to extract skeleton data from 50 videos. And each action has 7500 skeletal data points in this dataset.

IV. Result and Discussion

This session will discuss the results of our study.

Fig. 3. Test results for the images to child-adult identification model

A. Child-Adult Identification Model

Fig.2 depicts the results of the child-adult identification model for the test photos, while Fig. 3 illustrates the results of the multi-person environment testing (child and adult). As you can see, the results of the tests are more accurate in identifying the child and adult. The training accuracy of the model is 85%, and the testing accuracy is 93%. This approach is a method of identifying the child and adult in the video to recognize the behaviours. Fig. 3 shows the child-adult identification in a multi-person environment. This approach used 1500 child and adult photos, dividing them into 1200 (80%) for training and 150 (10%) for validation and testing. An online annotation tool called makesense.ai is used to annotate images manually. (https://www.makesense.ai/) make sense is an open-source, web-based free annotation application.

B. *Comparing Action recognition models*

This approach used the ANN, 1DCNN, LSTM, and GNN (STGCN) to model the action recognition of the child and adult. We aim to explore their effectiveness in this session.

TABLE I. COMPARISON OF ADULT'S ACTION RECOGNITION MODEL ACCURACIES

Method	Accuracy		
	Training	*Validation*	*Testing*
ANN	90.23%	91.30%	91.35%
1DCNN	93.25%	93.30%	93.32%
LSTM	93.15%	93.20%	93.18%
GNN	95%	94.26%	94.36%

TABLE I displays the accuracies of the various models we employed for action recognition. The table shows that the ANN action recognition model has lower accuracy than 7500 skeletons, whereas GNN techniques have higher accuracy. As previously discussed, action detection using artificial neural networks fails to capture the interaction between frames. Thus, we turn to other 1DCNN, LSTM, and STGCN.

TABLE II. COMPARISON OF CHILD'S ACTION RECOGNITION MODEL ACCURACIES

Method	Accuracy		
	Training	*Validation*	*Testing*
ANN	93.12%	92.20%	92.31%
1DCNN	93.5%	93.2%	93.2%
LSTM	94.2%	93.5%	93.5%
GNN	96%	95.32%	94.20%

TABLE II shows the accuracy of the child action recognition model after using 1500 skeletal data and transfer-learning. The accuracy of this model is higher than that of the adult action recognition model. This model is possible due to the small dataset. For testing, we used different data from youtube. Here are some of the video results for each child and adult's actions.

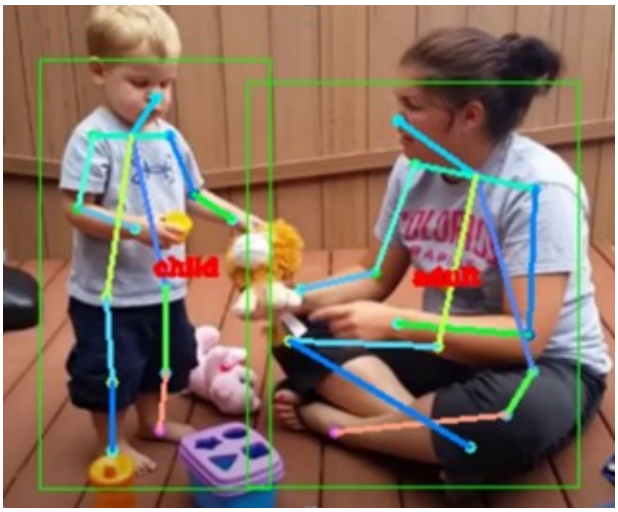

Fig. 4. Child and Adult Identification

Fig. 4 depicts the initial identification of a child and an adult, whereas Fig. 5 depicts recognizing the child's and adult's activities. We first distinguish between the objects (child or adult) and each skeletal sequence of the child and adult, which we then input separately into the two action recognition models.

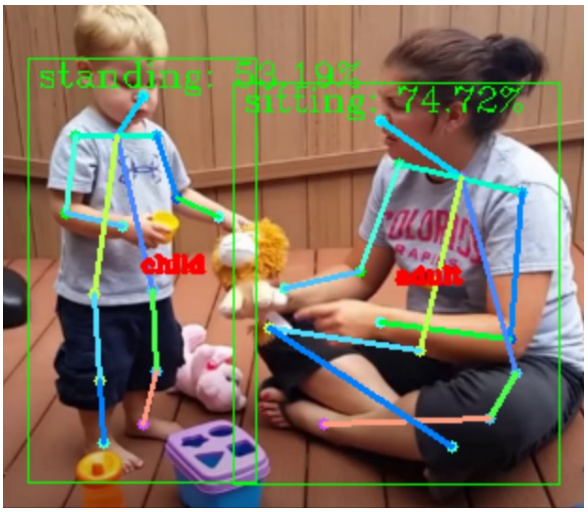

Fig. 5. Child's and Adult's action recognition

The percentage which belongs to the score in action is how effectively identify the key points in the skeleton. The excel sheet for the training used our skeleton dataset. As a result, 1DCNN can analyze this type of one-dimensional data. 2DCNN can operate with image data. We adjust the kernel size to get the temporal relationship. Furthermore, the GNN approach can determine the temporal relationship between the joints. We can sequence the joints in the frames.

TABLE III. COMPARATIVE ANALYSIS OF MULTI-PERSON ACTION RECOGNITION WITH STATE OF ARTS

Method	Action classes	Data	Testing Accuracy
3DConvNet [8]	12	Video	90.79%
LSTM [10]	16	Sensor	91.5%
Proposed Approach	6	Skeleton	94.36%

Table III shows the accuracy compared with some state of arts multi-person action recognition. This comparison is only based on adult actions because previous methods are based on adult action recognition. 3DConvNet based method used image data to identify the actions, and it achieved 90.79% average testing accuracy. LSTM based action recognition system comes with a 91.5% average testing accuracy. That approach used sensor data for identifying the actions. Actions based on gesture recognition. Our proposed method achieves 94.36% testing accuracy for skeleton data. We used the same video dataset and different videos for testing adults' actions. That's why some testing accuracy is higher than the training. The simulations are done using an NVIDIA GeForce RTX 2060 GPU and 64GB RAM.

V. CONCLUSION AND FUTURE WORKS

In automatic video analysis, action recognition has gotten a lot of attention, and it can help you save a lot of money on human resources for smart surveillance. Many methods are based on detecting only one action for a single person in a well-segmented video rather than recognizing several actions performed by multiple people at the same time in an untrimmed video. [22] This study focuses on recognizing child behaviours in a multi-person scenario while also recognizing adult activities. I must first identify the child in

the video to recognize the child's action. As a result, we created an object-detection model using yolov5 and an action recognition model based on the GNN (STGCN), which is more accurate than others. The temporal association between the frames in the video was not obtained using the artificial neural network method. As a result, this approach used the 1DCNN and, LSTM, GNN algorithms as a solution. However, the GNN approaches outperform the others in terms of training results.

I plan to train the recognition model using 3D skeletal data (including depth information) in the future. After that, we intend to create a model that includes action recognition and the object. As an example, the child is playing with a ball. The ball is another object, and it interacts with the child. Between the object detection model and the action detection model, there is a 30-frame delay because 30 frames are fed into an STGCN action recognition model to identify the actions. STGCN needs a set of frames to get the temporal relationship between the frames. I intend to reduce this delay in the future by comparing the accuracy of action recognition by feeding a lower number of frames to the model.

REFERENCES

[1] J. Dong, Y. Gao, L. H.J, H. Zhou, Y. Yao, Z. Fang, and B. Huang, "Action recognition based on the fusion of graph convolutional networks with high order features," Applied Sciences, vol. 10, no. 21.02.2020, p. 4, 2020.

[2] Pham, H. Hieu, H. Salmane, L. Khoudour, A. Crouzil, P. Zegers and S. A. Velastin, "Spatiotemporal image representation of 3D skeletal movements for view-invariant action recognition with deep convolutional neural networks," Sensors, vol. 19, no. 24.04.2019, 2019.

[3] Jegham, Imen, A. B. Khalifa, I. Alouani and M. A. Mahjoub, "Vision-based human action recognition: An overview and real-world challenges," Digital Investigation, vol. 32, 2020.

[4] Zhao, Rongyong, Y. Wang, P. Jia, C. Li, Y. Ma, and Z. Zhang, "Review of Human Gesture Recognition Based on Computer Vision Technology," in 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2021.

[5] Tsai, Jen-Kai, C.-C. Hsu, W.-Y. Wang and S.-K. Huang., "Deep learning-based real-time multiple-person action recognition system," Sensors, vol. 17, 2020.

[6] Wang, Limin, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. V. Gool, "Temporal segment networks: Towards good practices for deep action recognition," in European conference on computer vision, Springer, Cham, 2016.

[7] Mohtavipour, S. Mehdi, M. Saeidi, and A. Arabsorkhi, "A multi-stream CNN for deep violence detection in video sequences using handcrafted features," The Visual Computer (2021), pp. 1-16, 2021.

[8] Tsai, Jen-Kai, C.-C. Hsu, W.-Y. Wang and S.-K. Huang, "Deep learning-based real-time multiple-person action recognition system.," Sensors, vol. 17, 2020.

[9] Woo, Sungpil, J. Byun, S. Kim, H. M. Nguyen, and D. K. Janggwan Im, "RNN-based personalized activity recognition in the multi-person environment using RFID," in 2016 IEEE International Conference on Computer and Information Technology (CIT), 2016.

[10] Ordóñez, F. Javier and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," Sensors, vol. 16, p. 1, 2016.

[11] Yan, Sijie, Y. Xiong and D. Lin, "Spatial-temporal graph convolutional networks for skeleton-based action recognition," in Thirty-second AAAI conference on artificial intelligence, New Orleans, Louisiana, USA, 2018.

[12] I. OF, P. JS, S. J, and Y. BW, "Child and adult classification using the ratio of head and body heights in images," International Journal of Computer and Communication Engineering, vol. 3, pp. 120-122, 2014.

[13] "analyticsindiamag," .com, 19 12 2020. [Online]. Available: https://analyticsindiamag.com/. [Accessed 10 11 2021].

[14] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei and Y. Shekih, "OpenPose: real-time multi-person 2D pose estimation using Part Affinity Fields," IEEE transactions on pattern analysis and machine intelligence, vol. 43, pp. 172-186, 2019.

[15] A. D. Dongare, R. R. Kharde, and A. D. Kachare, "Introduction to an artificial neural network," International Journal of Engineering and Innovative Technology (IJEIT), vol. 2, pp. 189-194, 2012.

[16] Mumtaz, Wajid, and A. Qayyum, "A deep learning framework for automatic diagnosis of unipolar depression," International Journal of medical informatics, vol. 132, no. 01.12.2019, 2019.

[17] Ragab, M. G, S. J. Abdulkadir and N. Aziz, "Random search one dimensional CNN for human activity recognition.," in 2020 International Conference on Computational Intelligence (ICCI), 2020.

[18] A. S. and S. Juliet, "A Comprehensive Study on Human Activity Recognition.," in 2021 3rd International Conference on Signal Processing and Communication (ICPSC), 2021.

[19] C. F. P. Elias, J. Sedmidubsky and P. Zezula, "LSTM-based real-time action detection and prediction in human motion streams.," Multimedia Tools and Applications, vol. 78, pp. 27309-27331, 2019.

[20] Liu, Jun, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan and A. C. Kot, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," IEEE transactions on pattern analysis and machine intelligence, vol. 42, pp. 2684-2701, 2019.

[21] Kang, S. Min and R. P. Wildes, "Review of action recognition and detection methods," arXiv preprint arXiv, 2016.

[22] C. Gunagchun, Y. Wan, A. N. Saudagar and K. Namuduri, "Advances in human action recognition: A survey," arXiv preprint arXiv, 2015.