VIVA-Tech International Journal for Research and Innovation
ISSN(Online): 2581-7280

Volume 1, Issue 4 (2021)
Article No. X
PP XX-XX

VIVA Institute of Technology
9th National Conference on Role of Engineers in Nation Building – 2021 (NCRENB-2021)

VIVA-TECH IJRI
ANNUAL RESEARCH JOURNAL

VIVA-TECH INTERNATIONAL JOURNAL FOR RESEARCH AND INNOVATION
ANNUAL RESEARCH JOURNAL
ISSN(ONLINE): 2581-7280

Review of Pose Recognition Systems

Juilee Bhombe[1], Aashish Jethwa[2], Aditya Singh[3], Dr. Tatwadarshi Nagarhalli[4]
[1](Computer Engineering Department, Viva Institute of Technology, Mumbai, India)
[2](Computer Engineering Department, Viva Institute of Technology, Mumbai, India)
[3](Computer Engineering Department, Viva Institute of Technology, Mumbai, India)
[4](Computer Engineering Department, Viva Institute of Technology, Mumbai, India)

Abstract : *Human motion is fundamental to understanding behaviour. In spite of advancement on single image 3 Dimensional pose and estimation of shapes, current video-based state of the art methods unsuccessful to produce precise and motion of natural sequences due to inefficiency of ground-truth 3 Dimensional motion data for training. Recognition of Human action for programmed video surveillance applications is an interesting but forbidding task especially if the videos are captured in an unpleasant lighting environment. It is a Spatial-temporal feature-based correlation filter, for concurrent observation and identification of numerous human actions in a little-light environment. Estimated the presentation of a proposed filter with immense experimentation on night-time action datasets. Tentative results demonstrate the potency of the merging schemes for vigorous action recognition in a significantly low light environment.*

Keywords - *Action Recognition, artificial neural network, body part detection, computer vision, convolutional neural network, deep learning, deep neural network, human action recognition.*

1. INTRODUCTION

Statistics India said that in 2018, [9] there were 4,258,973 household theft cases, public fight cases, and border disputes. To confront the condition such as the cases mentioned above, generally, [1] a surveillance camera is used to watch properties and borders. Hence, if the theft and fight event happens, then the recorded data which is saved by the surveillance camera can be used as a reference to track the criminal action. However, the traditional surveillance camera is a passive device that cannot give any notification when a criminal event happens. In addition, there is a problem that the monitoring which is executed by using the surveillance camera in fewer lighting rooms makes the captured image not good enough. Various potential applications in diverse areas (e.g. optical surveillance, video renewal, sports video analysis, human-computer interfaces, and smart rooms) have encouraged the evolution of a big number of automated action recognition techniques [10]. These approaches consider many challenging scenarios like actions in the wild (YouTube videos), actions in the crowd, actions in group formations, actions in movies, actions across different viewpoints, and in the presence of occlusion [10]. However, the major approaches consider action recognition in high-quality day-time video sequences or in the presence of bright lighting conditions and do not focus on adverse lighting or the recognition of actions during night-time. Considering the case of a little-light place in an environment with sparse illumination. Any visual processing and manipulation of intensity values in such imagery result in different types of undesired artefacts like noise enlargement, intensity congestion, and dropping of resolution. The problem motivates the employ of numerous sensors often of complementary nature. A general many sensor night observation systems employ low light images by low light visible cameras and infra-red images by forwarding looking infra-red cameras.

2. LITERATURE SURVEY

2.1 Literature of existing system

Ching-Hang Chen and Deva Ramanan [1] have suggested a method for 3 Dimensional estimations of the pose from a sole RGB image. It has a two-step approach, firstly they have done 2D estimations using Deep Neural Nets. It produces accurate predictions even with self-occlusions. Then they have used memorization techniques to lift 2Dimensional poses to 3Dimensional. This is feasible because of the 3Dimensional MOCAP data accessibility. They have used a probabilistic formulation i.e. they have considered a joint probability. In which there are two terms the first term comes from a statistic neighbor-nearest (NN) model and the second term comes from image-based CNN that predicts 2 Dimensional key point heatmaps. So, the technology that they have used here is CNN and NN model. For manufacturing heatmaps, it has a pose machines model of convolutional which is trained on MPII dataset. The dataset it has used is Human3.6M. Just the NN model manufactures a standard fidelity of 70.93 and the overall mean accuracy in all the poses is 82.72. The paper suggested a method only for a single RGB frame. The two-step approach i.e. 2D + 3D and then again wrapping exemplar on camera gives accurate results. It plans on predicting the poses using the same model but for multiple frames and in non-ideal situations.
Li Shen, Ying Chen [7] have proposed human pose estimation algorithms using deep learning-based methods for mark less pose estimation from single depth maps are established on a framework of usual that lay hold of a depth map 2D and straight revert the convolutional neural networks (CNNs) 2 Dimensional nexus along 3 Dimensional coordinates of the body of a human. Nevertheless, the map's depth is essentially 3 Dimensional information, serve it as 2 Dimensional images will deform the figure of the genuine object through prediction from 3 to 2 Dimensional space, and compels the network to perform perspective distortion-invariant estimation Moreover, rightly reverting 3 Dimensional coordinates from a 2 Dimensional image is an extremely nonlinear plotting, which genesis trouble in the learning procedure. To defeat the complications, a module called Supervised En-decoder is proposed to process 3D convolution data, which can also be stacked through series connection to adapt different sizes of the dataset. Based on the module, a network called Supervised High Dimension En-decoder Network is designed, which has been handed down to predict key points of mark fewer humans in a sole map depth in 3D space. Trials show improvised prediction accuracy with the mean of 92% collates to the art-of-the-state proceed towards.

Widodo Setyo Yuwono, Dodi Wisaksono Sudiharto, Catur Wirawan Wijiutomo [9], human detection is generated by using Face Detection for the first mechanism and for the second mechanism is Head and Shoulders Detection. Both mechanisms are formed in master/slave.Thus, if the master function which is the Face Detection cannot recognize the object as a human, then the Head and Shoulders detection is going to be shown. The inability of master function for detection happens because of less illumination, or an obstacle such as the human face appears in the opposite direction from the lens of the camera. To minimize the detection failure, the night vision feature is also presented in this study for the surveillance camera prototyping.

The human object detection by using the Head and Shoulders method as a slave feature is still established for minimizing the failure probability. OpenCV to capture images or videos, the libraries of OpenCV can be enforced. OpenCV contains 2,500 algorithms that can be utilized to detect the human face, to recognize the object, to know the movement, to track the motion, etc. Telegram to notify the house owner, Telegram app can be applied by installing it in the Raspberry Pi. Telegram is a media social application that is free and is open source which provides chat and bot services. The proposed system works properly, especially when the system uses the combination of Face Detection and also Head and Shoulders method. The usage of the night vision feature also improves the accuracy for detecting the object moves more than 80% in several variations of illumination.

Anwaar Ulhaq [10] suggest concurrent recognition of action from video streams of multiple utilizing deep manifold-vision depiction learning. Moreover, it initiates a spatial-temporal feature-based filter of correlation, for synchronous observation and identification of numerous human actions in little-light conditions. It has estimated the production of the suggested filter with sizeable experimentation on night-time action datasets. Trial results indicate the effectiveness of deep fusion schemes for robust action recognition in extremely low-light conditions. Night Vision Action Dataset (NV): The dataset contains video sequences recorded by using two separate cameras: Raytheon Thermal IR-2000B and Panasonic WVCP47.

This dataset represents a good collection of human actions performed in daytime and night-time settings with different sets of challenges for recognizing actions. It has divided all datasets into coaching and checking videos and utilizing70% of videos for coaching and 30% for checking. It is observed that their filter performance for the NV action dataset outperforms other competitive approaches. It has achieved an overall average precision of

80.3%. It achieved low performance with an average precision of 75.0% and 71.1% for punch and handclasp actions.

Muhammed Kocabas, Nikos Athanasiou, Michael J. Black [11], Black [11], it suggested Video Inference for Body Pose and Shape Estimation (VIBE), which makes use of a current scale of huge capture of motion dataset (AMASS) jointly with leftover, in the wild, key point annotations of 2D. The key newness is a disruptive framework of learning that grasps AMASS to discriminate between actual motions of humans and those formed by their secular pose and shape regression networks. And explained the architecture of a novel temporal network with the attention of self-mechanism and displays that disruptive training, at the sequence level, yields kinematically reasonable motion sequences without in the wild ground truth 3 Dimensional labels. Executed substantial examination to scrutinize the essentiality of movement and explain the potency of VIBE on challenging 3D pose estimation datasets, achieving the art of the stage performance. It imprecise model of the body of SMPL variables for each mount in a sequence of video using a temporal generation of network, which is trained jointly with a movement differentiator. The differentiator has access to a huge corpus of motions of humans in SMPL format. It has extracted the features of each frame using a pre-trained Convolutional Neural Network.

It is trained on a terrestrial encoder formed of aligned Gated Recurrent Units (GRU) that outputs inactive parameters having data intrinsic from past and future frames. Later, the characteristics are used to revert the variables of the SMPL body model at each time instance. Evaluation of state-of-the-art models on 3DPW, MPI-INF-3DHP, and Human3.6M datasets. VIBE (direct comp.) is the suggested model coached on video datasets similar to, while it is coached with extra information from the 3DPW set of coaching. VIBE outruns all art of the state models together with SPIN on the challenging in-the-wild datasets (3DPW and MPI-INF-3DHP) and obtains comparable results on Human3.6M. They have achieved accuracy in between 82% and 87%.

Lei Wang, Du Q. Huynh, and Piotr Koniusz [12] have juxtaposed the 10 latest Kinect-based recognition of action algorithms. The Kinect camera is used for video-based action recognition. The performance is reliant on the type of characteristics being extricated and the actions are represented. The writers have examined and algorithms were compared for both subject of cross recognition of action and cross-view of cross recognition of action using datasets of six namely, MSRAation3D, 3D Action Pairs, Cornell Activity Dataset, UWA3D Activity Dataset, UWA3D Multiple viewpoint Activity II and NTU RGB+D Dataset.

It summarizes a few parameters like skeleton-based features work well than depth based features then handcrafted features performed well than deep learning features for small datasets but deep learning methods achieved good results on large datasets. In spite of quality precision in cross identification of subject action and very low precision in cross-view action recognition, there are various problems in these algorithms like different viewpoints, visual appearances, human body sizes, lighting conditions, partial occlusions, and self-occlusions.

Carlos Roig, Manuel Sarmiento, David Varas, Issey Masuda, Juan Carlos Riveiro and Elisenda Bou-Balust [13] propose a multi-modal human action recognition task in videos. The method takes into consideration action recognition, scene understanding, object detection, and acoustic event detection for human action recognition. All of this is done using a Pyramid Feature Combination architecture which comprises of two parts: a fusion block that performs domain-specific attention and combines pairs of features from different domains i.e. audio and video and a pyramid approach that uses the previous fusion block at different levels of hierarchy, resulting in a system that refines the input task-specific features and enriches features from other domains for human action recognition.

The pipeline used for human action detection uses a Yolov2 trained with MSCOCO. The Yolov2 network has been chosen over a Faster R-CNN for computational performance reasons. The dataset used here is a subset of the Moments in Time dataset. According to the experiment performed by the authors, there is an increase in accuracy from 24.97% using action features, to 35.43% with the full pyramid. This method effectively extracted features because it takes into consideration various factors while featuring extraction i.e. action, object, scene, audio.

Sungjoo Park, Dongchil Kim [14] have developed Recognition of detailed human actions in harsh environments such as low light environments is an issue that needs to be constantly resolved in a video surveillance system. To reduce this issue we can use 3D depth maps and get information about the human pose in harsh environments. In the paper, it suggested well-planned action recognition using 3 Dimensional images based on convolution neural

networks (CNN). As well as applied it to the video surveillance system which is intelligent and measures the detection accuracy by the action. Tentative results display that the action recognition accuracy for security is 61.5%. To achieve this accuracy they have combined MSD Action3D and D dataset + NTU RGB as well as made the own KETI dataset with 13 classes and 1000 samples.

Yahui Zhao; Ping Shi; Jian You [15], Since the human action recognition categories are increasing over the years the traditional supervised learning model has become increasingly difficult to collect enough training data to identify all categories. for some well-trained traditional supervised learning models, but it is a waste of time to collect enough samples of new categories and retrain them together in order to identify new categories. Proposes a mapping between visual features of video and semantic description of fine-grained human action recognition. Unlike most current zero-shot learning methods, which use manual features as visual features, we see features learned from I3D network models as visual features, which are more general than manual features. The accuracy of random guess in this technique is 1/10 but when we combine nearest + self-learning it becomes 28.2±0.02.

Peng Wang ; Yuliang Yang; Wanchong Li; Linhao Zhang ; Mengyuan Wang; Xiaobo Zhang; Mengyu Zhu[16] they have proposed a human action recognition (HAR) method based on convolutional neural networks (CNN), and used for a mortal flash of motion recognition. Earliest, accumulating data in scenarios of three and Deep Convolutional Generative Adversarial Networks (DCGAN) is accustomed to register enhancement of data to devise the dataset (DataSR). Subsequently, the 3*3 and 1*1 convolution kernels are used to outline the full convolution network and the model is further squeezed using the group convolution to obtain HARNET- the new model. Trials demonstrate that the map of HARNET is 94.36% of the DataSR dataset, and 76M is the model size, which the size of the YOLOv3 model is 30%. The DCGAN network is used to enhance the semaphore action dataset, and 3000 semaphore action data is input to train in DCGAN to generate a new dataset.

Using the network to create novel pictures is not 100% reliable, but it guarantees the difference between the images and is beneficial to training in the network. Model coaching and checking were performed using the processed semaphore action data, and the influence of different factors on the presentation of the model was examined. The paper uses the open-source Darknet framework for experimental research, the environment configuration is as follows: Ubuntu16.04, cuda8.0, cudnn5.0, GPU (TIAN XP), python 2.7.8, 12 GB memory. Model parameter settings: batch is 64, subdivisions are 32, learning_rate is 0.001, and max_iters is 20000. The convolution used in the design of the paper is a standard convolution, then tried to replace it with a deep separable convolution. Under the premise of maintaining accuracy, the model is further compressed, so that the designed deep learning algorithm can be applied in embedded platforms or edge computing products.

Amir Nadeem; Ahmad Jalal; Kibum Kim [17], it has used scrutiny of a straightforward hallmark for the creation of characteristics from the parts of the body identified. The priority aim of the research is to merge analysis of straightforward hallmark with an Artificial Neural Network (ANN) for right observation and identification of human action. The suggested mechanism detects intricate actions of humans in two states of the datasets of art, i.e. KTH-dataset and Weizmann Action of Human. They have acquired attributes of multidimensional from twelve body parts and calculated from models of the body. The multidimensional characteristics are used as an insert for the artificial neural network (ANN). To ingress the efficacy of the suggested method, it has contrasted the end results with other artists of the state classifiers.

Experimental results show that the suggested technique is authentic and suitable in health exercise systems, intelligent surveillance, e-learning, not normal behavioral identification, safeguarding for infant misuse, intelligent image retrieval, and human-computer interaction. It uses Forefront disunion via skin tone is achieved using heuristic thresholding which is a method of picture transform operation. It identifies skin regions with aid of a suitable YCC model. Conversion from the RGB color model to the model of YCC is embellished in the suggested paper. In the basics of body core-points, five body parts are identified which are hands, feet, and head. Humans are detected from silhouette as given in the Algorithm. A total of twenty-five subjects has achieved these actions in different scenes, having a 25fps of frame rate. The recognition accuracy mean is 87.57% for the dataset

Analysis Table:
The table shows the analysis of survey of the existing system by stating the paper title, Accuracy and Data Used, Frame/Person, Architecture Components and Detection Environment for 3D Pose Estimation.

Table 2.1 Analysis for survey of existing system

Sr. no	**Paper Title**	**Accuracy and Dataset Used**	**Frame/ Person**	**Architecture Components**	**Detection Environment**
1	3D Human Pose Estimation = 2D Pose Estimation + Matching [1]	Human3.6M gives 82.72	Single RGB frame/ single person	CNN and NN model	Indoors
2	HEMlets Pose: Learning Part-Centric Heatmap Triplets for Accurate 3D Human Pose Estimation [2]	20% more with Human3.6M. 15.2% more with HumanEva-I and 38.0% more with MPI-INF-3DHP	Single frame/ single person	Image feature extractor ResNet-50 then a ConvNet for image feature upsampling then again a ConvNet is used for HEMlets learning and 2D joint detection. Finally, a pose regression module for 3D adopting a soft-argmax operation for pose estimation of 3D Outdoor	Outdoor
3	Integral Human Pose Regression [3]	Human3.6M gives preciseness of: 78.7,, MPII gives accuracy: 87% and COCO gives accuracy: 74%	NA	Deep convolutional backbone network to extract convolutional target output from the features	NA
4	LCR-Net++: Multi-Person 2D and 3D Pose Detection in Natural Images [4]	MuPoTs-3D dataset reports 74% accuracy	Single frame/ Multiple people	Localization-Classification-Regression architecture	Natural Images
5	Deep 3D Human Pose Estimation under Partial Body Presence [5]	Human3.6M gives 88.00%	Single frame/ single person	Deep CNN to regress the human pose.	Outdoor
6	Adversarially Parameterized Optimization for 3D Human Pose Estimation [6]	Human3.6M gives 92.00%	Single frame/ single person	Supervised En-decoder is proposed to process 3D convolution data, which can also be stacked through a series connection to adapt different sizes of datasets.	Outdoor
7	Supervised High-Dimension Endecoder Net: 3D	62.00%	Single frame/	The system uses 3D-PAFs to infer 3D limb vectors and combine	Outdoor

	End to End Prediction Network for Mark-less Human Pose Estimation from Single Depth Map [7]		Multiple persons	them with 2D poses to regress the 3D coordinates	
8	Design and Implementation of Human Detection Feature on Surveillance Embedded IP[9]	No dataset used as it is hardware-based implementation. The NN model produces an accuracy of 70.93 and overall accuracy in all the poses is 82.72.	Single frame/ Multiple people	OpenCV to capture images or videos, the libraries of OpenCV can be enforced. OpenCV contains 2,500 algorithms that can be utilized to detect the human face, to recognize the object, to know the movement, to track the motion, etc.	Real-time photos outdoor
9	Action Recognition in the Dark via Deep Representation Learning[10]	Night Vision Action Dataset (NV)dataset contains two separate cameras: Raytheon Thermal IR-2000B and Panasonic WVCP47	Single frame/ Multiple people	framework for action recognition based on deeply fused convolutional features.use Spatio-temporal features, C3D features as building blocks of our framework.	Videos of actions like a punch, push, etc.
10	VIBE: Video Inference for Human Body Pose and Shape Estimation[11]	large-scale motion capture dataset (AMASS) wild datasets (3DPW and MPI-INF-3DHP) and obtains comparable results on Human3.6M.	Single frame/ multiple people	novel temporal network architecture with a self-attention mechanism.	Videos outdoor

This table analyses the survey of existing systems by mentioning the Paper title, Accuracy, Dataset Used, Model/Architecture used for action recognition mechanism.

Table 2.1 Analysis for survey of existing system

Sr. no.	**Paper Title**	**Accuracy**	**Dataset Used**	**Model/ Architecture used**
1.	A Comparative Review of Recent Kinect-Based Action Recognition Algorithms [12]	Not giving satisfactory results.	MSRAation3D, 3D Action Pairs, Cornell Activity Dataset, UWA3D Activity Dataset, UWA3D Multiview Activity II, and NTU RGB+D.	Kinect Cameras

2.	Multi-Modal Pyramid Feature Combination for Human Action Recognition [13]	Increase in accuracy from 24.97% using action features, to 35.43% with the full pyramid	MSCOCO	Pyramid Feature Combination architecture
3.	Study on 3D Action Recognition Based on Deep Neural Network [14]	61.5%	KETI dataset with 13 classes and 1000 samples.	3D depth maps with CNN
4.	Fine-grained Human Action Recognition Based on Zero-Shot Learning [15]	The accuracy of random guess in this technique is 1/10 but when we combine Nearest + self-learning it becomes 28.2±0.02.	No dataset is given	I3D network models as visual features
5.	Research on Human Action Recognition Based on Convolutional Neural Network [16]	The mAP of HARNET on the DataSR dataset is 94.36%, and the model size is 76M, which is 30% of the size of the YOLOv3 model.	Generation of dataset DataSr	new model HARNET , Deep Convolutional Generative Adversarial Networks (DCGAN) and open-source Darknet framework
6.	Human Actions Tracking and Recognition Based on Body Parts Detection via Artificial Neural Network [17]	87.57% is the recognition accuracy of mean for the dataset.	The two state of the art datasets, i.e. KTH-dataset and Weizmann Human Action.	YCC model is illustrated

In this above mentioned papers, all the experimental results from the experiments outlined in table 2.1 are presented and examined in detail. Any of the above-stated papers does not have good accuracy and efficacy. There is no system to provide night light recognition. Methods discussed in the papers are only for a single person recognition not for multi-person recognition. All of the implementations of the above paper works on single frame recognition instead of multiple frames. No system does pose estimation on natural or real images or on real-time videos. Night vision detection is not present. There is confusion between detecting some poses like punch, push and slap. On detecting undesirable activity there is no module to notify those actions.

3. CONCLUSION

The relentless surge in the crime at day, as well as night time, has led the country's banal people to be the victim of miscreants who attack or sometimes kill the person by looting them or by asking ransom and these scoundrels never get caught on camera or by actions they do. So, to create a 3D pose estimation system that recognizes the poses of multiple people in multiple frames in a real- time video throughout the day as well as night. Any of the above-mentioned papers does not have good accuracy and precision. There is no system to provide night light recognition. Methods discussed in the above papers are only for a single person recognition not for multi-person recognition. All of the implementations of the above paper works on single frame recognition instead of multiple frames. No system does pose estimation on natural or real images or on real- time videos.

Acknowledgements

We would like to express a deep sense of gratitude towards Computer Engineering Department for their constant encouragement and valuable suggestions. The work that we are able to present is possible because of her/his

timely guidance. We would like to pay gratitude to the panel of examiners Prof. Sunita Naik, Dr. Tatwadarshi P. N., Prof. Dnyaneshwar Bhabad, & Prof. Vinit Raut for their time, effort they put to evaluate our work and their valuable suggestions time to time. We would like to thank Project Head of the Computer Engineering Department, Prof. Janhavi Sangoi for her support and coordination. We would like to thank Head of the Computer Engineering Department, Prof. Ashwini Save for her support and coordination. We are also grateful to the teaching and non-teaching staff of the Computer Engineering Department who lend their helping hands in providing continuous support.

REFERENCES

[1] Ching-Hang Chen and Deva Ramanan. 3D Human Pose Estimation = 2D Pose Estimation + Matching. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018.

[2] Kun Zhou, Xiaoguang Han, Nianjuan Jiang, Kui Jia and Jiangbo Lu. HEMlets Pose: Learning Part-Centric Heatmap Triplets for Accurate 3D Human Pose Estimation. In IEEE/CVF Conference on Computer Vision (ICCV), 2019.

[3] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang and Yichen Wei. Integral Human Pose Regression. In CPVR, 2018.

[4] Grégory Rogez, Philippe Weinzaepfel and Cordelia Schmid. LCR-Net++: Multi-Person 2D and 3D Pose Detection in Natural Images. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[5] Saeid Vosoughi and Maria A. Amer. Deep 3D Human Pose Estimation under Partial Body Presence. 25th IEEE International Conference on Image Processing (ICIP), 2018.

[6] Dominic Jack, Frederic Maire, Anders Eriksson, Sareh Shirazi. Adversarially Parameterized Optimization for 3D Human Pose Estimation. 2017 International Conference on 3D Vision (3DV)

[7] Li Shen, Ying Chen. Supervised High-Dimension Endecoder Net: 3D End to End Prediction Network for Mark-less Human Pose Estimation from Single Depth Map. 2019 5th International Conference on Control, Automation and Robotics (ICCAR).

[8] Ding Liu, Zixu Zhao, Xinchao Wang, Yuxiao Hu, Lei Zhang, Thomas Huang. 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). Improving 3D Human Pose Estimation Via 3D Part Affinity Fields.

[9] W. S. Yuwono, D. Wisaksono Sudiharto and C. W. Wijiutomo, "Design and Implementation of Human Detection Feature on Surveillance Embedded IP Camera," *2018 International Conference on Sustainable Information Engineering and Technology (SIET)*, Malang, Indonesia, 2018, pp. 42-47, doi: 10.1109/SIET.2018.8693180.

[10] A. Ulhaq, "Action Recognition in the Dark via Deep Representation Learning," *2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS)*, Sophia Antipolis, France, 2018, pp. 131-136, doi: 10.1109/IPAS.2018.8708903.

[11] M. Kocabas, N. Athanasiou and M. J. Black, "VIBE: Video Inference for Human Body Pose and Shape Estimation," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 5252-5262, doi: 10.1109/CVPR42600.2020.00530.

[12] Lei Wang, Du Q. Huynh and Piotr Koniusz. A Comparative Review of Recent Kinect-Based Action Recognition Algorithms. In IEEE Transactions on Image Processing (Volume: 29), 2019.

[13] Carlos Roig, Manuel Sarmiento, David Varas, Issey Masuda, Juan Carlos Riveiro and Elisenda Bou-Balust. Multi-Modal Pyramid Feature Combination for Human Action Recognition. In IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019.

[14] Sungjoo Park, Dongchil Kim. Study on 3D Action Recognition Based on Deep Neural Network. 2019 International Conference on Electronics, Information, and Communication (ICEIC).

[15] Yahui Zhao, Ping Shi, Jian You. Fine-grained Human Action Recognition Based on Zero- Shot Learning. 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS).

[16] P. Wang *et al*., "Research on Human Action Recognition Based on Convolutional Neural Network," *2019 28th Wireless and Optical Communications Conference (WOCC)*, Beijing, China, 2019, pp. 1-5, doi: 10.1109/WOCC.2019.8770575.

[17] A. Nadeem, A. Jalal and K. Kim, "Human Actions Tracking and Recognition Based on Body Parts Detection via Artificial Neural