Human Action Recognition Using Deep Data: A Fine-Grained Study

D. Surendra Rao*
*Research Scholar,
Koneru Lakshmaiah Education Foundation, India
Associate Professor,
Guru Nanak Institutions Technical Campus
India
sdustakar@gmail.com

Sudharsana Rao Potturu[a]
[a]Associate Professor
Koneru Lakshmaiah Education Foundation
India
potturusd54@gmail.com

Bhagyaraju. V[b]
[b]Professor and Principal
Siddhartha Institute of Engineering and Technology
India
vbhagya01@gmail.com

Abstract: The video-assisted human action recognition [1] field is one of the most active ones in computer vision research. Since the depth data [2] obtained by Kinect cameras has more benefits than traditional RGB data, research on human action detection has recently increased because of the Kinect camera. We conducted a systematic study of strategies for recognizing human activity based on deep data in this article. All methods are grouped into deep map tactics and skeleton tactics. A comparison of some of the more traditional strategies is also covered. We then examined the specifics of different depth behavior databases and provided a straightforward distinction between them. We address the advantages and disadvantages of depth and skeleton-based techniques in this discussion.

Keywords: *Depth, action recognition, depth maps, skeleton, feature extraction, and classification.*

1. Introduction

Human Action Recognition (HAR) has grown in importance in computer vision in recent years, and it has made considerable strides over the last decade. Furthermore, HAR [1] is gaining traction in several fields, including Human-Computer Interaction (HCI) [2], telemedicine, automation, assistive living, video retrieval, and digital surveillance. The primary goal of HAR is to automatically interpret and classify the processing activities of an undisclosed video. Several forms of apps benefit from activity identification from recordings. For example, an optical surveillance device [3] with an automated action recognition system can aid in the prevention of robberies at public locations such as airports, metro stations, and bus stops.

A tremendous amount of research been conducted to obtain high-level knowledge of human behavior. HAR is described as detecting the behavior of objects or actors present in the data for an input image or series of images. Based on their sophistication, we split human attitudes into four categories: "Gestures," "Acts," "Interactions," and "Group Activities." [3] Human body parts such as the head and fingers have straightforward gestures [4]. Next, movement can be defined as a collection of gestures that includes more than one gesture, such as tossing, walking, hand-clapping, and so on. On the other hand, encounters are human behavior involving at least two human beings or things. Interactions include handshakes between two people, basketball shooting, tennis serving, and so on. Finally, there are more individuals involved in the social events. As an example, a group of people marching down the street, playing cricket, or taking part in another activity.

The primary role in the HAR phase is to keep track of an actor's actions in real-time. This can be accomplished by obtaining specific data and categorizing them into two groups depending on the data used as HAR input. They are focused on vision and distance maps. The HAR models in the first group employ computer vision methods to analyze visual observations obtained from optical sensing devices such as cameras and infrared sensors [5]. Even though extensive testing has been done on the HAR focused on vision artifacts [6-8], they have several flaws. These approaches' general problems include a wide range of operations, scalability, reusability, etc. The computationally typical algorithms in signal processing and computer vision require large amounts of hardware to be feasible. The data-dependent on vision still lacks 3D knowledge [9].

Due to their ability to provide 3D data, inexpensive depth sensors (ex. Microsoft Kinect sensor) have led to widespread usage of HAR [10].The Kinect sensors have gotten a far and wide relevance in so many commercial games since they are modest and can remove the full-body movements from an overall client. The power of a

Manuscript received June 5, 2022
Manuscript revised June 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.6.16

depth sensor to collect both depth and color information simultaneously is its main benefit. Using depth cameras, action recognition system realism is improved and issues associated with RGB recordings are removed -as can be seen in Table 1, the advantages and disadvantages of RGB video cameras and depth cameras.

Table 1: Pros and Cons Comparison Between Different Cameras

	Depth Cameras	RGB video cameras
Pros	1. Insensitive to texture and color changes. 2. Easy to operate 3. Able to deliver a 3D structure of information 4. inexpensive and widely available 5. Not sensitive to lighting condition and illumination variations	1. inexpensive and widely available 2. ensures a rich texture information 3. easy to operate
Cons	1. No color information 2. Sensitive to the presence of different objects and materials in the FoV.	1. need the presence of actor or object in the field of view (FoV) 2. Computer vision algorithms are much more complex to implement 3. Sensitivity to the calibration of CAM. 4.Sensitivity to lighting conditions, illumination variations, and cluttered backgrounds

Due to these benefits of depth sensors, so many researchers have developed different types of HAR models by proposing different computational algorithms that consider the depth action videos as input. This paper outlines the earlier developed HAR models based on the depth action data. Initially, we explore the details of different depth action datasets and the generalized evaluation metrics. Next, we explore a detailed survey over the HAR methods that focused on analyzing depth information. Under this depth information, we have considered both depth maps and posture data, and finally, a simple comparison is outlined at the end of the paper.

Accordingly, the remainder of the paper is organized as follows; section 2 explains the information of different kinds of depth action datasets. Section 3 explores the complete details of the state of the art survey, and finally, the conclusions are provided in section 4.

2. Depth action datasets

The development of many depth action datasets has aided research on HAR using RGB-D sensors. These datasets are generated using Microsoft Kinect sensors, a unique sensor (called depth sensors). The descriptions of the most often used datasets are seen in Table.2. Most of the datasets deliver the essential information captured with the help of the RDG-D device, i.e., depth and color frames and posture data. In the posture data, the action video is represented in spatial coordinates. The usual range of actions of all these datasets is observed as 10 to 20, and the average number of subjects or actors employed to construct is approximately 10. Instead of capturing each action only once, most of the datasets acquired the actions after making the actors carry out each action 2 to 3 iterations. The primary purpose behind the construction of these datasets is twofold; (1) Human-Computer Interaction and (2) Daily Activity (DA). The datasets focused on the HCI-based applications may consist of a *sidekick, draw a circle, and draw across. They are generally acquired* in a simple background even though they are most challenging due to the similar characteristics in many gestures.

On the other hand, the datasets that focused on the DA may include drinking, eating, walking, running, etc. Further, in some datasets, they were acquired from real-time scenarios due to which some kind of occlusions will appear and have complex backgrounds. The figure shows examples of RGB and depth images as well as skeleton action samples collected from different datasets.



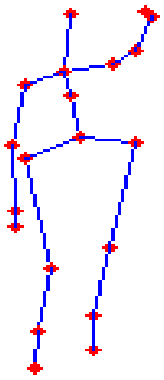

Fig.1 Samples of different actions data formats (a) RGB, (b) Depth and (c) Skeleton

Table 2: List of Datasets That Are Focused On The Depth Information

Name	Year Acquired	Number of Actors used	Number of actions	Application type	Data format	Number of times	Number of samples
NTURGBD+ [11]	2016	40	60	DA/HCI	RGB, Depth, Skeleton	-	56880
UTD-MHAD [12]	2015	8	27	HCI	RGB, Depth, Skeleton	4	861
KARD [13]	2014	10	18	DA/HCI	RGB, Depth, Skeleton	3	540
UPCV Action [14]	2014	20	10	DA	Skeleton	-	-
3D Online Action [15]	2014	24	7	DA	RGB, Depth, Skeleton	-	-
IAS-Lab Action [16]	2013	12	15	DA	RGB, Depth, Skeleton	3	540
WorkoutSu-10 Gesture [17]	2013	15	10	DA	Depth, Skeleton	10	1500
Berkeley MHAD [18]	2013	12	11	HCI	RGB, Depth	5	660
CAD-120 [19]	2013	4	10	DA	RGB, Depth, Skeleton	-	120
MSR Action Pairs [20]	2013	10	6	DA	Depth	3	180
LIRIS Human Activities [21]	2012	21	10	DA	RGB, Depth	-	49
ACT4 Dataset [22]	2012	24	14	DA	RGB, Depth	1	6844
Florence 3D action [23]	2012	10	9	DA	RB, Skeleton	2 or 3	215
MSR Daily Activity [24]	2012	10	16	DA	RGB, Depth, Skeleton	2	320
G3D [25]	2012	10	20	HCI	RGB, Depth, Skeleton	3	-
UTKinect Action [26]	2012	10	10	DA/HCI	RGB, Depth, Skeleton	2 or 3	200
MSR Gesture 3D [27]	2012	10	12	HCI	Depth	2 or 3	336
CAD-60 [28]	2012	4	12	DA	RGB, Depth, Skeleton	-	60
DHA [29]	2012	21	23	DA/HCI	Depth	-	483
MSR Action 3D [30]	2010	10	20	HCI	Depth, Skeleton	2 or 3	567

3. Literature Survey

The general model of HAR is employed in three phases: feature extraction, dimensionality reduction, and classification [4]. The model tries to extract the features from input action data in the feature extraction phases. These features describe the action present in the input action image or video in a compact representation such that the system can understand. Next, in the dimensionality reduction phase, the size of the feature vector is reduced, and finally, in the classification phase, a classifier is modeled to classify the actions. Among the three steps, feature extraction is the most important, and most researchers only concentrated on this aspect. As a result, we divide the current approaches into categories depending on the function used. We also differentiated the approaches based on the data type they have considered due to the sophisticated quality of depth sensor data. (1) The two types of approaches are (1) depth map-based approaches and (2) skeleton-based approaches.Table.3 presents a list of various behavior recognition processes.

3.1 Depth map-based approaches

Methods based on depth maps take depth maps as input and extract either global or local features from a space-time volume. Comparatively to visual data, depth maps provide action information that is not affected by lighting. However, designing efficient and effective depth map-based representations to recognize actions is challenging. A few examples of depth maps can be seen in Figure 2.

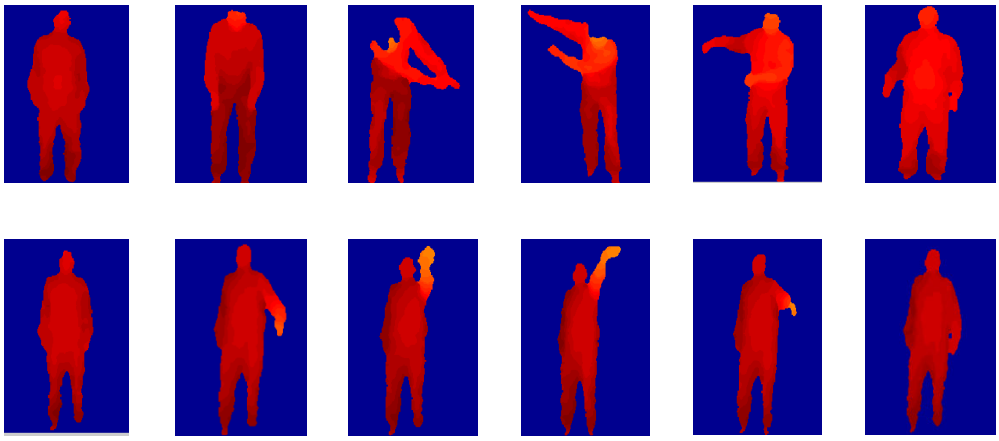

Fig.2 Images of golf-swing maps (above) and high waves (below) from the MSR action 3D dataset

Li et al. [30] studied human action recognition from depth maps. In this method, the authors employed encoding the actions in the expandable graphical model through bag-of-points to formulate an action graph [31]. Every node of this graph describes a salient posture represented through a small set of 3D points sampled from depth maps. The shape of the action is described by 3D points and the statistical distribution is described by Gaussian Mixture Models (GMM). An experiment is conducted on the MSR Action 3D data set.

Nevertheless, [30] has one major drawback: it loses spatial context information. In addition, because the occlusions are different from side to top angles, the reliability of the actions is less. Due to this problem, sampling interest points becomes very difficult for other persons' given actions. Space-Time Occupancy Patterns (STOPs) were developed by Vieira et al. [32], a new action descriptor to address the problem. In depth action videos, a 4D time-space grid is presented. Based on a saturation scheme, the human silhouette's moving parts were emphasized by boosting the positions of sparse cells. Experiments are conducted.

Next, Wang et al. [33] focused on the issue of noise and occlusion in-depth charts, and to solve it, they turned the 3D action series into a 4D form and added a pattern called the Random Occupancy Pattern (ROP). ROPs are derived from 4D sub-volumes that are uniformly sampled at various positions and sizes. Since this approach removes characteristics wider, it is less susceptible to noise and occlusions. In addition, to improve recognition accuracy, this approach used a sparse coding scheme [34] and weighted random sampling. MSR Action 3D dataset is used to carry out

the experiments, and the results are comparable to those of [30] and [31].

Jalal A et al. [35] used depth silhouettes and R-transform [36] to characterize the action based on the performance of outlines in action identification. Random Transform is used in the feature extraction process to render the device scale and translation invariant for a given depth silhouette. After Principal Component Analysis (PCA) and Linear Discriminate Analysis (LDA) have been used for dimensionality reduction, one prominent action descriptor is extracted using LDA. As a final step, a Hidden Markov Model is used to classify actions.

Through Depth Motion Maps (DMM), Yang et al. [37] have developed a novel action descriptor that captures the temporal energies of sequences. This method involves projecting depth maps onto orthogonal planes and transforming the projected images into normalized images. Each map is given a binary map dependent on thresholding, then added together to form DMM. The Histogram of Directed Gradients (HOG) is computed for each DMM. SVM algorithm is employed for classification. The significant advantage of DMM is its less complexity when compared with other methods.Using the MSR Action 3D dataset, experiments are conducted.

In a similar way, Chen et al. [38] extracted features from depth action videos using Local Binary Patterns (LBPs) after projecting the video from three views: front, rear, and top. This is achieved using the Kernel-Based Extreme Learning Machine (KELM) [49]. Feature-based fusion, and judgment-based fusion, have also proven to be useful components of this technique at the fusion stage. When the LBP features are merged at the function point, the softmax rule is used to combine the classification scores at the decision phase. Furthermore, the DMMs were created by the same source, Chen et al. [39], at the section stage, where the depth video sequence is segmented into multiple overlapping segments. After that, DMM is used to classify each element, and LBP extracts position rotation invariant information. Fisher kernel generates a compact function vector for each operation in the final level. For action grouping, ELM is used.

M. A1-Faris et al. [40] introduced a newfound variant of DMM called "Fuzzy weighted multi-resolution DMMs (FWMDMMs)" that focuses on segmentation and motion detail. By segmenting the temporal action frames at various stages, this model centered on developing multiple DMMs at multiple levels. The weight function was used in three orientations for finding the meaning after the DMMs were represented: linear, reverse, and central. Finally, the FWMDMMs are fed into a deep CNN model for classification.A new approach, referred to as MFSS (Multilevel Frame Select Sampling) has been proposed by Xu Weiyao et al. [42] to produce three stages of temporal samples based on the depth of the input sequence.Secondly, they are represented using Motion and Static Mapping (MSM), Block-based & LBP representations, and Fisher kernel representations. For action classification, KELM has been effective. Wu Li et al. [43] extended the LBP to "Discriminative Completed LBP (DiscLBP)." Two classifiers were proposed: collaborative representation classification (CRC) and DMM-assisted behavior identification (DMM).

It was proposed by Kim D et al. [44] to represent depth action maps compactly. To generate the side view, the depth action picture from the front view is first processed. Following that, two additional descriptors, "Depth Motion Appearance (DMA)" and "Depth Motion History (DMH)," is used to characterize both the side and front views of the action picture. For action labeling, an SVM classifier is used. Finally, the operation is defined using HOG. On the other hand, prior depth charts did not consider the various motions of body pieces.

Using depth maps and Local Gradient Auto-correlations (GLAC) [47], Chen et al. introduce yet another HAR framework. DMMs of depth action images were used for this method to derive shift-invariant image features. GLAC's main accomplishment is the capability to generate 2nd-order gradients which are capable of exploring the rich information about edge features. The final step in action classification is the use of ELM, a single hidden layer feed-forward neural network [48], after concatenating GLAC and DMM features. A new feature extraction method called Space-time Autocorrelation of Gradients (STCOG) in 3D space was introduced by Chen et al. [50] in order to improve the further recognition performance. In this method, initially, the DMMs are computed to transform the depth image into shape and motion cues. The next step is to extract features based on auto-correlation data of image local gradients [51], which can compensate for the loss of temporal information in DMMs.

Liu H et al. [52] came up with another method to deal with the loss of temporal information in DMMs. As described in [53], Hierarchical Depth Motion Maps (HDMM) are used for feature extraction and classification using Convolutional Neural Networks (CNNs). The novelty of this approach is to create the

mimics of an action image in the view of camera rotations. Secondly, HDMM can be used to extract the body's shape and movement at different time scales.The 3 CNNs are used for three views (front, side, and top) projected onto the orthogonal planes.

As part of the new HON4D action descriptor [54], Prefer ad Liu leverages depth maps to create an action descriptor. HON4D describes the action of obtaining the surface normal orientation distribution in a 4D volume of spatial coordinates, time, and depth using Histograms. To create the HON4D, the 4D space is first quantized using vertices from a regular polychoric. The quantization is then used to find the most discriminative and dense area using a novel discriminatory density measure. Data from the MSR Action 3D dataset is used to conduct experiments.

The authors of Zhang et al. [55] proposed to represent human action using 4D spatial-temporal features along with the depth maps. The 4D feature vector is generated by combining the geometric and visual components weighted linearly. This is done by concatenating the per-pixel responses with gradients contained within the Spatio-temporal window. K-means are used to cluster the function vectors for dimensionality reduction. They used the Latent Dirichlet Allocation (LDA) [57] model to forecast events, and Gibbs sampling [56] was used for preliminary estimation and inference. Based on 198 short video clips of six types of actions, they validated the algorithm with a self-collected dataset.

3.2 Skeleton Based Approaches

The real inspiration for the skeleton-based activity recognition was initiated by Johansson [58], which demonstrated that the alone joint positions could recognize the large set of actions. Unlike the methods that focused on the depth data, most of the skeleton-based approaches explicitly model the temporal dynamics. The fundamental explanation for this significance is their natural skeleton correspondence over time, which was difficult to achieve with depth-based results. There are three methods to obtain skeleton data in general: (1) Single view depth maps, (2) Multi-view color images, and (3) Active Motion Capture (MoCAP) [59, 60]. The significant difference between the skeleton data acquired through these models is embedded noise. Compared to the first two models, the skeleton data obtained through MoCAP is cleaner. When correlated to the remaining two, the multi-view setup is generally employed to acquire color images, and thus, they produce skeletons those have more stability than the monocular depth maps. In earlier, most of the action recognition methods were employed over the multi-view skeleton data and MoCAP data while the recent works focused on the skeleton data from monocular depth maps, because the setup is very simple. In the following section, we discuss different action recognition methods those are developed based on skeleton data. Figure.3 shows some examples of skeleton images.

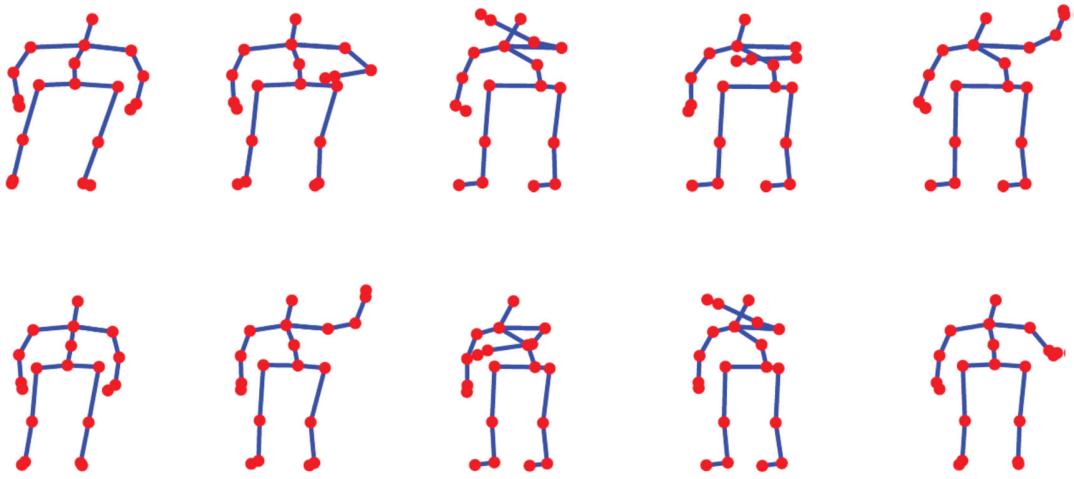

Fig.3 examples of skeleton action images from MSR Action 3D dataset, draw tick (top) and draw cross (bottom)

As a result of projecting the 3D joint trajectory onto low-dimensional space, Campbell [61] and Bobick [62] represented human motion as curves. This phase space is defined independently of the body's position in relation to each axis. In the phase space, an action is

represented with a curve while the static is represented as a single point. An action feature is represented after the projection of curves those are learned from multiple 2D spaces through supervised learning. But, due to the consideration of an action as a simple curve and curve fitting problem, only few types of actions those have simple movements are only recognized. Since phase space representation is invariant across scales and views, they are scale and view invariant.

Histogram of 3D Joint Locations (HOJ3D) proposed as a new feature descriptor by Xia et al. [26] to consider Hip center as a root and encode the spatial occupancy information.They proposed a new method in which the hip center serves as an origin point for a new version of a spherical coordinate system, segmenting the 3D space into multiple bins. However, this method is not focused on the computation of radial distance which makes the system not scale invariant. Unlike the Actionlet ensemble [24], for the estimation of spatial occupancy, this method employed probabilistic voting. LDA is adopted for dimensionality reduction, Vector Quantization is adopted for normalization and discrete HMM is used to model action dynamics followed by action recognition. For experimental validation, they have considered their own dataset along with MSR Action 3D dataset. However, the heavy dependency on the hip center may affect the recognition accuracy when the actor is not facing towards camera. In Jinag et al. 's paper [62], they propose Skelton context to be the invariant to the absolute body orientation and position. A multi-scale pairwise position distribution for each joint in the skeleton is extracted to quantify correspondence between postures. Bag-of-words be evaluated by the Conditional Random Fields (CRFs). On the other hand, [26, 62] assumes that the original joints of the skeleton run parallel to the ground.

Due to the lack of motion hierarchy, the above-mentioned methods are having limitation for multiple human actions. To solve this issue, Koppula et al. [19, 63] mainly focused on the Human-object interactions. They used the markov random field (MRF) to analyse the action video chain, which had two types of nodes: sub-activity nodes and object nodes, with the edges between nodes representing the interaction between objects and sub-activities. Both types of nodes have their own set of characteristics**.** The feature defied for objects is oriented to the object's location and its displacements within a temporal segment. These are tracked through SIFT tracker. Next, the feature of sub-activity is evaluated from the skeleton data acquired from skeleton tracker on RGBD video. For experimental validation, they have considered their own dataset called as Cornell 120 along with Cornell 60 dataset [64].

Sung et al. [28, 65] also employed the action hierarchy model of a maximum Entropy Markov Model of two layers (MEMM). In this method, two types of nodes are defined, one for the representation of sub-activities and another for the representation of complex activities. Every action is analyzed using four different features, namely (1) body posture converted into a local coordinate system (2) hand positions relative to the head and torso (3) joint motion with a temporal sliding window and (4) images and point clouds with HyperObject Geometry. GMM is used for action recognition and for the experiments; they have used their own dataset.

Along with point cloud information, Wang et al. [24] used the skeleton information. There is a rule of thumb that some actions will vary when they are performed in the presence of an object, and in that case just using the skeleton data will not suffice. To solve this issue, they have introduced a novel actionlet ensemble which captures the intra-class variances through Local Occupancy pattern (LOP). On the basis of the 3D point cloud surrounding a given joint, LOP features are evaluated. In order to derive the Fourier Temporal Pyramid features at every joint, the authors concatenated both features and then used Short Fourier Transform. The MSR Action 3D dataset, the CMU MoCap dataset, and the MSR Daily Activity dataset were used as experimental validation data for action classification.

The skeleton motion was encoded by Yao et al. [66], using the geometric relationship between particular joints to describe the skeleton motion. For action recognition, they have employed Hough Forest [68]. Moreover the experiments are conducted on a multi-view kitchen dataset [69]. X. Yang et al. [70] combined joint position differences with Eigen joints to form a new feature descriptor.By using this descriptor, we are able to describe the offset, motion, and static posture of the body joints.

The motion and posture features encode the temporal and spatial configurations with pairwise joint difference in a single frame and also between following frames, respectively The offset characteristics are then used to depict the difference between a pose and the first pose. For multi-class grouping, the "Nave-Bayes-Nearest-Neighbor (NBNN)" algorithm is used, and the MSR Action 3D dataset is used for simulation.

SMIJ is defined by F.Ofli et al. [71] as a series of K representative poses made up from the skeleton frames. The skeleton of joints is exceptionally interpretable estimates like most extreme angular velocity of joints, variance or mean of joint angles etc. At that point, the activity succession is address with Histogram of its posture words. For exploratory approval, they have utilized various datasets. In addition, the activity acknowledgement technique proposed by M. Barnachon et al. [72] additionally centered around bunching yet the grouping is cultivated through Hausdroff distance. The middle component of each bunch is considered as pose of cluster. Nonetheless, the histogram-based action representation techniques ignore the temporal information in favor of only using statistical data.

Through translations and rotations in 3D space, R. Vemulapalli et al. [73] evaluated geometric relationships between different body parts to model the skeleton action. The skeleton joints are represented by curves in lie group using the proposed representation. Then the feature from curve space is transformed into the lie algebra and then performed classification through linear SVM, Fourier pyramid temporal representation and dynamic time wrapping (DTW) [79]. For experimental validation, they have used three datasets

Researchers presented 3D skeleton joint trajectories to support 3D action recognition [74]. Each skeleton is first converted into three clips which each contain different frames and are then used to train deep neural networks. Clips originate from one cylinder and frames originate from one frame, where each frame represents the temporal data of the whole skeleton sequence. In order for this to be achieved, Convolutional Neural Networks (CNN) were used to learn the long-term temporal information, followed by Multi-task Learning to synthesize all the information. The experimenters have used three datasets for experimental validation: the NTURGBD+ dataset, the SBU Kinect interaction dataset [76], and the CMU dataset [75].

A new model used to model dynamic skeletons is referred to as Spatial-Temporal Graph Convolutional Networks (ST-GCN) by Yan et al. [77]. ST-GCN is more advanced than the conventional skeleton based methods and it ensures an automatic learning of both temporal and spatial patterns from action data. This method has a stronger generalization capability. For experimental validation, they have used two datasets; they are NTURGBD+ dataset, and Kinects [78].

A. Kamel et al. [80] proposed an action fusion method by combining depth maps and postures through CNNs. For both actions inputs, two different action descriptors are derived. To attain increased recognition accuracy, three different CNN channels are involved and the outputs obtained at each channel are fused. Further several fusion scores are employed to analyze the effect of different fusion rules.Three datasets have been used for simulation purposes: 1) MSRAction3D; 2) Multimodal Texas at Dallas; and 3) Multimodal Action Dataset (MAD).

L. Cai et al. [81] used depth sequence features and CNN for HAR. Initially the DMM are extracted from depth sequence after the projection into three Cartesian planes. To further accelerate the computation and also reduce complexity, a two-dimensional process identification and 3-D input architecture are proposed. Simulations are performed on the basis of MSR Action 3D, UT-Kinect, and a private CTP action 3D dataset.

An enhanced spatial-temporal graph convolution network (MS-ESTGCN) based on multi-streaming has been proposed by Li et al. [82]. For the aggregation of temporal features, each block of MS-ESTGCN is employed for Graphic Convolutional Layers (GCLs) with different kernel sizes. For simulation purpose, two dataset namely NTU-RGBCD and Kinetics-Skeleton are employed.

Y. Han t al. [83] proposed a GL-LSTM+ Diff model for 3D HAR. The Global Spatial Attention (GSA) model provides precise information about the movements in human actions by representing the weights for different kernels. Moreover, accumulative learning curves are implemented to enhance the frames with the highest contribution. For classification, they have employed LSTM based RNN.Researchers conduct rigorous experiments on SBU's common small dataset as well as NTU's largest RGBCD dataset.

Table 3: Comparison Of Different Action Recognition Methods Based On Depth Data

Reference	Taxonomy	Representation	Classifier	Dataset for simulation	Year
Li et al. [30]	Depth	Bag-of-3D points and 2D projection	Action graph	MSR Action 3D	2010
Vieira et al. [32]	Depth	STOPs followed by PCA	Action graph	MSR Action 3D	2012
Wang et al.	Depth	ROP and Sparse Coding	SVM	MSR Action 3D	2012

[33]					
Jalal A et al. [35]	Depth	Depth silhouette and R-transform	HMM	Self-created Daily activity dataset	2012
Yang et al. [37]	Depth	DMM followed by HOG	SVM	MSR Action 3D	2012
Chen et al. [38]	Depth	DMM in three views (front, side and top) followed by LBP	KELM	MSR Action 3D	2015
Chen et al. [39]	Depth	Segmented DMM followed y LBP and fisher kernel	ELM	MSR Action 3D	2016
M. A1-faris et al. [40]	Depth	Segmentation, FWMDMMs	CNN	MSR 3D daily action and MSR 3D actions ,Northwestern-UCLA multi-view action 3D,	2019
Xu Weiyao et al. [42]	Depth	MSM followed by LBP and fisher kernel	KELM	MSR Gesture 3D, and UTD-MHAD, MSR Action 3D,	2019
Wu Li et al. [43]	Depth	DMM followed by DiscLBP	ELM and CRC	MSR Action 3D	2018
Kim D et al. [44]	Depth	DMA, DMH and HOG	SVM	MSR Action 3D	2014
Chen et al. [46]	Depth	DMMs and GLAC	ELM	MSR Gesture 3D,MSR Action 3D	2015
Chen et al. [50]	Depth	DMMs and STCOG	ELM	MSR Gesture 3D,MSR Action 3D	2016
Liu H et al. [52]	Depth	HDMMs	CNN	MSRAction3D and DHA	2017
Oreifej ad Liu [54]	Depth	HON4D	SVM	MSR Gesture 3D, and MSR Daily Activity 3D,MSR Actions 3D,	2013
Xia et al. [26]	Skeleton	HOJ3D and LDA	HMM	MSR Action 3D	2012
Jinag et al. [62]	Skeleton	Bag of words and CRF			2015
Koppula et al. [19, 63]	Skeleton	Pose and object features	Multi-class SVM	Cornel-120 and Cornel-60 action datasets	2012
Sung et al. [28, 65]	Skeleton	HOG, Pose features, and GMM	MEMM	Self-created dataset	2011
Yao et al. [66]	Skeleton	Geometrical relational features	Hough Forest	Multi-view kitchen	2012
X. Yang et al. [70]	Skeleton	Offset, motion and static features of body joins	NBNN	MSR Action 3D	2014
F. Ofli et al. [71]	Skeleton	SMIJ and Histograms	Levenshtein distance	MSR Action 3D, Self-created dataset and HMDB05	2014
R. Vemulapalli et al. [73]	Skeleton	Geometric relationships between different body parts, DTW and Fourier pyramid temporal features.	SVM	MSR Action 3D, UTKinect-Action and Florence3D-Action	2014
Q. Ke et al. [74]	Skeleton	3D trajectories of skeleton joints	CNN	NTURGBD+, SBU Kinect interaction and CMU	2017
S. Yan et al. [77]	Skeleton	ST-GCN	CNN	NTURGBD+ and Kinect	2018
A. Kamel et al. [80]	Depth and Skeleton	Depth Motion Maps and Skeleton joint descriptors	CNN	MSRAction3D; and i) multimodal action dataset (MAD) dataset, ii) University of Texas at Dallas-multimodal human action dataset;	2019
L. Cai et al. [81]	Depth	DMM on three planes	3-D CNN	MSR Action 3D, UT-Kinect and a private CTP action 3D dataset	2018
Fanjia Li et al. [82]	Skeleton	Skeleton with GCL	CNN	NTU-RGBCD and Kinetics-Skeleton	2020
Y. Han t al. [83]	Skeleton	global spatial Attention (GSA)	RNN	NTU RGBCD and SBU dataset	2020

4. Discussion and Conclusion

In the recent years, the depth data has been acquiring a huge research interest in different kind of applications. Among those applications, the major contribution is done in the field of human action recognition. Compared to the tradition vision based data, the depth data has more advantages and the development of applications based on depth data has more benefits. For example, the RGB images are sensitive to illumination and lighting conditions while the depth data is insensitive. Based on this inspiration, so many authors put effort towards the improvement of various types of HAR models. Based on this comparison we understood that the depth data has so many advantages tan the vision data like less computational cost, more efficiency, inexpensiveness and ease of deployment.

Our study explores different approaches of action recognition based on depth. Initially we have outlined the basic prospects of depth data and explored its advantages by comparing with RGB data. Further we have outlined a detailed description about the depth action datasets those are used in the performance evaluation of different HAR methods. Among the surveyed datasets, there are two types of datasets, they are the dataset those are intended mainly on the recognition of daily activities and the second one is the datasets those are intended on provision of interaction between human and computer. Making the computer to behave like a human being is a tedious task and only a limited number of interactions are acquired in the HCI datasets. Further, we also noticed that in the depth datasets, broadly there are two kinds of data formats, they are one is depth maps and another is skeleton. These datasets are typically categorized into three types of formats: RGB, depth, and skeleton.

Next, we have conducted a detailed survey on the HAR methods. Broadly all the surveyed methods are categorized into two categories; These approaches are skeleton-based and based on depth data. In the old category, the developed HAR system seeks the depth maps as input while in the second one it seeks the skeleton data as input. The initial methods like ROP, HOG, STOP, and HON4D etc. applied the traditional action representation techniques like STIPs, cuboids and occupancy patterns, bag of words etc. as feature extraction techniques which are not new ones. In contrast to these methods, DMM was introduced first by Yang et al. [37] who showed it significantly increased recognition accuracy, as well as computing complexity. Compared to the traditional action representation methods, the DMM is very simple and also effective. Based on these advantages, so many versions of DMMs like HDMM, FWMDMM, DMA, DMH etc. are developed and proven the effectiveness of DMM. However, the major issue with DMM its non-robustness for noises, occlusions and some small side effects like minor movements and body shaking movements. For a given action sequence with these disturbances the DMM and its subsequent methods had shown a limited performance. Hence there is a need to work on such kind of issues in future to achieve an efficient recognition performance.

Based on skeleton data, the next category of methods represents the movement by showing the joints' positions in three dimensions (x-, y-, and z-axis). At the starting phase methods, the action is described by the computation of displacements of joints in successive frames. However, they are susceptible for view point variations. Hence to achieve view-point variance, some authors transformed the skeleton joints data in Cartesian plane to spherical plane and the actions are described through the radial distance and angular deviation with horizontal and vertical axis. However, they are observed to have susceptibility to minor actions or the actions with similar movements like Draw cross, Draw tick in MSR action 3D dataset. We conclude from these observations that there is still a great deal of potential for further research in the field of depth data assisted HAR.

References

[1] I. Theodorakopoulos, D. Kastaniotis, G. Economou, S. Fotopoulos, "Pose-based human action recognition via sparse representation in dissimilarity space", *Journal of Visual Communication and Image Representation,* 2014; 25(1):12-23.

[2] S. Sempena, N. U Maulidevi and P. R. Aryan, "Human action recognition using Dynamic Time Warping", *Proc. of the 2011 International Conference on Electrical Engineering and Informatics*, Bandung, Indonesia, 2011.

[3] Chen Chen, Roozbeh Jafari and Nasser Kehtarnavaz, "A survey of depth and inertial sensor fusion for human action recognition", *Multimed Tools Appl*, 76, 4405-4425, 2017.

[4] Aggarwal JK, Xia L., "Human activity recognition from 3d data: a review", *Pattern Recognition Letters*, 48:70–80, 2014.

[5] Klette, R., Tee, G., "Understanding human motion: A historic review", In Rosenhahn, B., Klette, R., Metaxas, D., eds.: Human Motion. Volume 36 of Computational Imaging and Vision. Springer Netherlands (2008) 1-22.

[6] Chen C, Kehtarnavaz N, Jafari R, "A medication adherence monitoring system for pill bottles based on a wearable inertial sensor*", 36th IEEE Annual International Conference on Engineering in Medicine and Biology Society (EMBC)*, 2014, pp. 4983–4986.

[7] Shah M, Javed O, Shafique K. "Automated visual surveillance in realistic scenarios", *IEEE Multimedia*, 2007; 14(1):30e9.

[4] Aggarwal J, Ryoo M. "Human activity analysis", *ACM Comput Surv*, Jan. 2011; 43(3):1-43.

[8] Chen L, Khalil I., "Activity recognition: approaches, practices and trends", In: *Activity recognition in pervasive intelligent environments Atlantis ambient and pervasive intelligence*, vol. 4; 2011. p. 10-31.

[9] Michalis Vrigkas, Christophoros Nikou and Ioannis A. Kakadiaris, "A Review of Human Activity Recognition Methods", *Frontiers in Robotics and AI*, Volume 5, Article 28, 2015.

[10] Poppe R., "A survey on vision-based human action recognition", *Image Vis Comput*., 28(6), 2010, pp. 976–990.
[11] Ramanathan M, Yau WY, Teoh EK, "Human action recognition with video data: research and evaluation challenges", *IEEE Trans Human-Machine Systems*, 44(5):650–663, 2014.
[12] Aggarwal JK, Xia L., "Human activity recognition from 3d data: a review", *Pattern Recognition Letters*, 48:70–80, 2014.
[13] A. Shahroudy, J. Liu, T.-T. Ng, G. Wang. "NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR);* 2016. Las Vegas, NV; pp. 1010-1019.
[14] C. Chen, R. Jafari, N. Kehtarnavaz., "UTD-MHAD: A Multimodal Dataset for Human Action Recognition Utilizing a Depth Camera and a Wearable Inertial Sensor", *IEEE International Conference on Image Processing (ICIP);* 2015. Québec City, Canada; pp. 168-172.
[15] S. Gaglio, G. L. Re, M. Morana., "Human activity recognition process using 3-d posture data", *IEEE Transactions on Human-Machine Systems*, vol.45, issue 5, 2015, pp.586-597.
[16] G. Yu, Z. Liu, J. Yuan., "Discriminative Orderlet Mining for Real-Time Recognition of Human-Object Interaction", In: D. Cremers, I. Reid, H. Saito, M.-H. Yang, editors. Computer Vision – ACCV 2014: 12th Asian Conference on Computer Vision, Singapore, Singapore, November 1-5, 2014, Revised Selected Papers, Part V. Image Processing, Computer Vision, Pattern Recognition, and Graphics ed. Springer International Publishing, Cham; 2014. pp. 50-65.
[17] M. Munaro, G. Ballin, S. Michieletto, E. Menegatti., "3D flow estimation for human action recognition from colored point clouds", *Biologically Inspired Cognitive Architectures*. 2013;5:42-51.
[18] F. Negin, F. Özdemir, C. B. Akgül, K. A. Yüksel, A. Erçil., "A Decision Forest Based Feature Selection Framework for Action Recognition from RGB-Depth Cameras", In: M. Kamel, A. Campilho, editors. Image Analysis and Recognition. Lecture Notes in Computer Science ed. Munich: Springer, Berlin, Heidelberg; 2013. pp. 648-657.
[19] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, R. Bajcsy., "Berkeley MHAD: A Comprehensive Multimodal Human Action Database", *IEEE Workshop on Applications of Computer Vision (WACV*); 2013. Clearwater, Florida; pp. 53-60.
[20] H. S. Koppula, R. Gupta, A. Saxena., "Learning human activities and object affordances from RGB-D videos", *International Journal of Robotics Research,* 2013; 32 (8):915-970.
[21] O. Oreifej, Z. Liu., "HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences", *IEEE Conference on Computer Vision and Pattern Recognition*; Portland. 2013. pp. 716-723.
[22] Z. Cheng, L. Qin, Y. Ye, Q. Huang, Q. Tian. "Human Daily Action Analysis with Multi-view and Color-Depth Data", In: A. Fusiello, V. Murino, R. Cucchiara, editors. Computer Vision - ECCV 2012. Workshops and Demonstrations. Lecture Notes in Computer Science ed. Springer, Berlin, Heidelberg; 2012. pp. 52-61.
[23] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, P. Pala., "Recognizing Actions from Depth Cameras as Weakly Aligned Multi-part Bag-of-Poses", I*EEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW*); 2013. Portland, Oregon; pp. 479-485.
[24] J. Wang, Z. Liu, Y. Wu, J. Yuan., "Mining Actionlet Ensemble for Action Recognition with Depth Cameras", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR*); 2012. Providence, Rhode Island; pp. 1290-1297.
[25] V. Bloom, D. Makris, V. Argyriou., "G3D: A Gaming Action Dataset and Real Time Action Recognition Evaluation Framework", *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW*); 2012. Providence, Rhode Island; pp. 7-12.
[26] L. Xia, C.-C. Chen, J. Aggarwal., "View Invariant Human Action Recognition Using Histograms of 3d Joints", *IEEE Conference on Computer Vision and Pattern Recognition Workshops*; 2012. Providence, Rhode Island; pp. 20-27.
[27] A. Kurakin, Z. Zhang, Z. Liu., "A Real Time System for Dynamic Hand Gesture Recognition with a Depth Sensor", *In: Proceedings of the 20th European Signal Processing Conference (EUSIPCO*); 2012. Bucharest, Romania; pp. 1975-1979.
[28] J. Sung, C. Ponce, B. Selman, A. Saxena., "Unstructured Human Activity Detection from RGBD Images", *IEEE Conference on Robotics and Automation (ICRA)*; 2012. St. Paul, Minnesota; pp. 842-849.
[29] Y. C. Lin, M. C. Hu, W. H. Cheng, Y. H. Hsieh, and H. M. Chen, "Human action recognition and retrieval using sole depth information," in ACM MM, 2012, pp. 1053–1056.
[30] W. Li, Z. Zhang, Z. Liu., "Action Recognition Based on a Bag of 3d Points", *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW*). 2010. San Francisco, CA; pp. 9-14.
[31] Li, W., Zhang, Z., Liu, Z., "Expandable data-driven graphical modeling of human actions based on salient postures", *IEEE Transactions on Circuits and Systems for Video Technology,* 18(11) (2008) 1499-1510.
[32] Vieira, A., Nascimento, E., Oliveira, G., Liu, Z., Campos, M., "STOP: Space-time occupancy patterns for 3d action recognition from depth map sequences", *In: Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications,* (2012) 252-259.
[33] Wang J, Liu Z, Chorowski J, Chen Z, Wu Y., "Robust 3D action recognition with random occupancy patterns", *Computer Vision e ECCV 2012 Lecture Notes in Computer Science*; 2012. p. 872-885.
[34] Lee H, Battle A, Raina R, Ng AY., "Efficient sparse coding algorithms", *Proc. 19th Ann. Conf. Neural Information Processing Systems*; 2007. pp. 801-808.
[35] Jalal A, Uddin MZ, Kim JT, Kim T S., "Recognition of human home activities via depth silhouettes and R transformation for smart homes", *Indoor Built Environ*, 2012; 21(1):184e90.
[36] Wang Y, Huang K, Tan T., "Human activity recognition based on R transform", *IEEE Conference on Computer Vision and Pattern Recognitio*n; 2007.
[37]. Yang, X., Zhang, C., Tian, Y., "Recognizing actions using depth motion maps based histograms of oriented gradients", *In: ACM International Conference on Multimedia,* (2012) 1057-1060.
[38] Chen, C., Jafari, R., & Kehtarnavaz, N., "Action recognition from depth sequences using depth motion maps-based local binary patterns". In Proc., of *2015 IEEE winter conference on Applications of computer vision (WACV),* 2015, pp. 1092–1099.
[39] Chen, C., Liu, M., Zhang, B. , Han, J. , Jiang, J. , & Liu, H., "3d action recognition using multi-temporal depth motion maps and fisher vector", In *IJCAI ,* 2016, (pp. 3331–3337).
[40] Mahmoud Al-Faris, John Chiverton, Yanyan Yang 2 and David Ndzi, "Deep Learning of Fuzzy Weighted Multi-Resolution Depth Motion Maps with Spatial Feature Fusion for Action Recognition", *J. Imaging* 2019, 5, 82; doi:10.3390/jimaging5100082.
[41] Jiang Li; Xiaojuan Ban; Guang Yang; Yitong Li; Yu Wang, "Real-time human action recognition using depth motion maps and convolutional neural networks", *International Journal of High Performance Computing and Networking*, 2019 Vol.13 No.3, pp.312 – 320
[42] Xu Weiyao, Wu Muqing, Zhao Min, Liu Yifeng, Lv Bo, and Xia Ting, "Human Action Recognition Using Multilevel Depth Motion Maps", *IEEE Access*, Volume 7, 2019, pp. 41811- 41822.
[43] Wu Li, Q. Wang, and Y. Wang, "Action Recognition Based on Depth Motion Map and Hybrid Classifier", *mathematical problems in engineering*, Vol.2018, Article ID 8780105, 10 pages.
[44] Kim D, Yun W. H, Yoon H. S, and Jaehong H. S, "Action recognition with depth maps using hog descriptors of multi-view motion*," in proc., of 8th International Conference on Mobile Ubiquitous Computing, Systems, Services, and Technologies*, UBICOMM, pp. 2308–4278, 2014.
[46] Chen C, Hou Z, Zhang B, Jiang J, Yang Y., "Gradient local autocorrelations and extreme learning machine for depth-based activity recognition", *Advances in Visual Computing Lecture Notes in Computer Science*; 2015. pp. 613-623.
[47] Kobayashi T, Otsu N., "Image feature extraction using gradient local autocorrelations", *Lecture Notes in Computer Science Computer Vision e ECCV 2008*; 2008. p. 346-358.
[48] Huang G-B, Zhu Q-Y, Siew C. K., "Extreme learning machine: theory and applications", *Neurocomputing* 2006; 70(1e3):489 - 501.
[49] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics B*, vol. 42, no. 2, pp. 513–529, 2012.
[50] Chen C, Zhang B, Hou Z, Jiang J, Liu M, Yang Y., "Action recognition from depth sequences using weighted fusion of 2D and 3D autocorrelation of gradients features", *Multimed Tool Appl,* 2016;76(3): 4651-69.
[51] Kobayashi T, Otsu N., "Motion recognition using local auto-correlation of space time gradients", *Pattern Recogn Lett,* 2012;33(9):1188e95.
[52] Liu H, He Q, Liu M., "Human action recognition using adaptive hierarchical depth motion maps and Gabor filter", *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP*), 2017.
[53] Wang P, Li W, Gao Z, Zhang J, Tang C, Ogunbona P., "Deep convolutional neural networks for action recognition using depth map sequences", arXiv preprint arXiv:1501.04686; 2015.
[54] Oreifej, O., Liu, Z., "HON4D: Histogram of oriented 4d normals for activity recognition from depth sequences", *IEEE Conference on Computer Vision and Pattern Recognition,* (2013)
[55]. Zhang, H., Parker, L., "4-dimensional local Spatio-temporal features for human activity recognition", In: *International Conference on Intelligent Robots and Systems.* (2011) 2044{2049
[56] Grifiths, T.L., Steyvers, M., Finding scientific topics. Proceedings of the National Academy of Sciences of the United States of America 101(Suppl 1) (2004) 5228-5235.

[57] Pritchard, J. K.; Stephens, M.; Donnelly, P. (June 2000), "Inference of population structure using multi-locus genotype data", *Genetics*. 155 (2): *pp.* 945–959.
[58] Johansson, G., "Visual motion perception", *Scientific American* (1975).
[59] Ye, M., Wang, X., Yang, R., Ren, L., Pollefeys, M., "Accurate 3d pose estimation from a single depth image", *IEEE International Conference on Computer Vision*. (2011) 731-738
[60] Criminisi, A., Shotton, J., Robertson, D., Konukoglu, E., "Regression forests for efficient anatomy detection and localization in ct studies", *Workshop on Medical Computer Vision*. (2010)
[61] Campbell, L., Bobick, A., "Recognition of human body motion using phase space constraints", *IEEE International Conference on Computer Vision*. (1995) 624-630.
[62] M. Jiang , J. Kong , G. Bebis , H. Huo, "Informative joints based human action recognition using skeleton contexts", *Signal Process. Image Commun*., 33 (2015) 29–40.
[63] Koppula, H.S., Gupta, R., Saxena, A.: "Human activity learning using object affordances from RGB-D videos", CoRR abs/1208.0967 (2012)
[64] Lai, K., Bo, L., Ren, X., Fox, D., "Sparse distance learning for object recognition combining RGB and depth information", *International Conferences on Robotics and Automation*. (2011) 4007-4013.
[65] Sung, J., Ponce, C., Selman, B., Saxena, A., "Human activity detection from RGBD images", *In: Plan, Activity, and Intent Recognition*. (2011)
[66] Yao, A., Gall, J., Van Gool, L., "Coupled action recognition and pose estimation from multiple views", *International Journal of Computer Vision,* 100(1) (2012) 16-37
[67] M uller, M., R oder, T., Clausen, M., "Efficient content-based retrieval of motion capture data", *ACM Transactions on Graphics,* 24 (2005) 677-685
[68] Gall, J., Yao, A., Razavi, N., Van Gool, L., Lempitsky, V., "Hough forests for object detection, tracking, and action recognition", *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2011)
[69] Tenorth, M., Bandouch, J., Beetz, M., "The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition", *IEEE Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences,* (2009).
[70] X. Yang and Y. L. Tian, "Eigen joints-based action recognition using naive-Bayes-nearest-neighbor," in *Computer vision and pattern recognition workshops* (CVPRW), 2012, pp. 14–19.
[71] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Sequence of the most informative joints (SMIJ): A new representation for human skeletal action recognition", *J. Vis. Commun. Image Represent.*, Vol. 25, no. 1, pp. 24-38, 2014.
[72] M. Barnachon, S. Bouakaz, B. Boufama, and E. Guillou, "Ongoing human action recognition with motion capture", *Pattern Recognit.*, vol. 47, no. 1, pp. 238-247, 2014.
[73] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group," CVPR, pp. 588–595, 2014.
[74] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A New Representation of Skeleton Sequences for 3D Action Recognition," *in CVPR*, June 2017
[75] F. Han, B. Reily, W. Hoff, and H. Zhang, "space-time representation of people based on 3d skeletal data: a review", arXiv preprint arXiv:1601.01006, 2016.
[76] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras., "Two-person interaction detection using body pose features and multiple instance learning", *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW*), pages 28–35, 2012.
[77] S. Yan, Y. Xiong, and D. Lin, "Spatial Temporal Graph Convolutional Net works for Skeleton-Based Action Recognition," *in AAAI*, 2018.
[78] Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. 2017. The kinetics human action video dataset. In arXiv:1705.06950.
[79] C. Wolf, J. Mille, E. Lombardi, O. Celiktutan, M. Jiu, M. Baccouche, E. Dellandréa, C.-E. Bichot, C. Garcia, B. Sankur., "The LIRIS Human Activities Dataset and the ICPR 2012 Human Activities Recognition and Localization Competition", In: LIRIS Laboratory, Tech. Rep. RR-LIRIS-2012-004, March 2012
[80] Aouaidjia Kamel, Bin Sheng, Yang Po, Ping Li, and Ruimin Shen, "Deep Convolutional Neural Networks for Human Action Recognition Using Depth Maps and Postures", *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, Volume: 49 , Issue: 9 , Sept. 2019, pp.1806-1819.
[81] C. Linqin, L. Xiaolin, F. Chen, and M. Xiang, "Robust Human Action recognition based on Depth Motion Maps and improved Convolutional Neural Networks", *Journal of Electronic Imaging*, Vol. 27, No.5, 2018.
[82] Fanjia Li, Aichun Zhu, Yonggang Xu, Ran Cui, And Gang Hua, "Multi-Stream and Enhanced Spatial-Temporal Graph Convolution Network for Skeleton-Based Action Recognition", IEEE Access, Volume 8, 2020, pp. 97757-97770.
[83] Yun Han, Sheng Luen Chung, Qiang Xiao, Wei You Lin, and Shun Feng Su, "Global Spatio-Temporal Attention for Action Recognition Based on 3D Human Skeleton Data", *IEEE Access*, Volume 8, 2020, pp.88604-88616.

Mr. D.Surendra Rao is a research scholar in ECE Department at KL University Hyderabad, India. He completed his B.Tech. in ECE and M.Tech in VLSI System Design from JBREC, Hyderabad, India. Mr Surendra is currently working as an Associate Professor at Guru Nanak Institutions Technical Campus, Hyderabad, India. He has more than 16 years of teaching experience. His Area of interest is Artificial Intelligence, Deep Learning, and VLSI Design.

Dr. Sudharsana Rao Potturu has received a B.Tech degree in Electronics and Instrumentation Engineering from JNTU Kakinada, A.P, India, in 2010, an M.Tech degree in Electrical Engineering from IIT Kharagpur, West Bengal, India, in 2013, and a Ph.D. degree from the Department of Electrical Engineering, IIT Roorkee, Uttarakhand, India, in 2019. From 2013 to 2014, he worked as an assistant professor at JNTU Kakinada. Presently he has been working as Associate Professor at KL University Hyderabad since 2020. His area of interest includes a control system, Artificial Intelligence, Machine Learning, model order reduction, biomedical engineering, and controller design.

Dr. V.Bhagya Raju completed his B.E. from Vasavi College of Engineering and M.Tech in Wireless & Mobile Communications, and Ph.D. from JNTU-H. Currently, he is working as a Professor and Principal in ECE Department at Siddhartha Institute of Engineering and Technology, Hyderabad, India.