A Survey on Human Action Recognition Using Depth Maps and Skeleton Postures

Vinoda Gopampallikar[1], Shashi Kant Gupta [2]

[1] Department of Computer Science & Engineering **Lincoln University College, Malaysia**
Department of CSE (AI&ML) **CMR Technical Campus, Hyderabad, Telangana** 501401
email id : pdf.vinodaresearch@lincoln.edu.my, vinodaresearch@gmail.com, **vinodareddy.cse@cmrtc.ac.in**

[2] Department of Computer Science & Engineering, **Lincoln University College, Malaysia**
email id : **raj2008enator@gmail.com, shashigupta@lincoln.edu.my**

Abstract This paper presents an extensive literature survey on human action recognition (HAR) methods utilizing depth data in combination with deep learning techniques. Under depth data–based approaches, we review methods that employ both depth maps and skeleton joints or body postures for action representation. Furthermore, in the feature extraction phase, we analyze works based on handcrafted features as well as those leveraging deep learning–driven feature extraction. The survey not only categorizes and discusses these methods but also evaluates their strengths and limitations, offering a comprehensive perspective on the current state of HAR research and highlighting key trends and future directions in depth-based action recognition.

Keywords: Human Action Recognition, Depth maps, Skeleton Joints, Handcrafted features , Deep Learning.

---

1. Introduction

Nowadays, the utilization of Surveillance Cameras (also known as Closed Circuit Television or CCTV) is being increased to monitor the private and public spaces throughout the globe. Law enforcement authorities and Governments have used Video surveillance in several instances ranging from investigation of crimes, traffic control, protection of government buildings and urban environments, monitoring the demonstrators in the context of criminal investigations. Since the visual surveillance is more beneficial in preventing the crime activities, some of the nation's spending huge expenditure towards its installation and execution. For example, Myanmar has launched a Visual surveillance project called a "Safe City" worth of 1.2 million dollars by installing 335 Huawei Artificial Intelligence (AI) Equipped Surveillance cameras in eight townships in the Capital, Naypyidaw. Further, the country planned to install the same system in Mandalay by the mid of 2021 followed by in the commercial capital, Yangon.

With the major aim of proving security from stealing, vandalism, terrorism and violence, Visual surveillance cameras are being installed in various places like city streets, shopping malls, banks, airports, metro stations, train stations etc. For instance, the London have approximately 41% of public places have CCTV equipment mounted with approximately 4,20,000 cameras. However, it is a big challenge for human operators to monitor all the available CCTV cameras. Hence, the video surveillance footage is often used to find the suspects of crime. To ease the visual surveillance and ensure a 24/7 monitoring followed by triggering an alert in typical situations, Human Action Recognition (HAR) is evolved as alternative solution. HAR is able to completely exploit the potential of Larger Visual Surveillance camera network and greatly improves the safety. Further, the HAR is regarded as a fundamental task in many visual surveillance applications. So, many researchers concentrated over it and suggested several methods.

A. Human Action Recognition

In most of Visual Surveillance applications, HAR is regarded as the Major Aspect which aims at interpretation of the ongoing visible human actions which have certain contextual meaning. Human action involves the movements in the whole body or only at some portions like Arms, Legs, Head, and Hands, etc. Based on the involvement of different parts of human action, the actions are categorized into four categories; they are 1) Gestures, 2) Actions, 3) Interactions and 4) Group Activities. Gestures are the basic movements of a person's body and they describe a meaningful motion of a person like *Raising leg* or *Stretching of an arm*. Next, actions are performed by single person and are considered as composition of multiple gestures. Actions are temporally organized Gestures and some examples are *Punching*, *Waving* and *Walking* etc. Interaction is formulated based on involvement of two or more persons and/or objects. For example a *person kicking the boxing bag* is called as human to object interaction and a fight between two persons is called as human to human interaction. Finally the

Group activity is formulated based on the involvement of multiple persons and/or objects like *two groups fighting, a meeting* and *agroup of person marching*

B. Motivation

Automatic recognition of human actions has great significance in real time applications. For example a robot with the action recognition capabilities can assist the patients in their physiotherapy activities. Similarly an autonomous vehicle with action recognition capabilities can avoid the pedestrians from dangerous accidents. However, the HAR has several application constraints means it shows efficiency in only few environments. For instance a HAR model developed with an objective of health care industry won't show an effective performance in other applications like Human computer interaction, video retrieval etc. From this analysis, we understand the requirements of HAR to get more prominent results and try to develop a robust and generalized model for human action recognition. Hence there is a necessity of a robust and generalized action recognition model which is the major motivation behind our research on HAR.

A huge research has been accomplished in past years over the development of an effective HAR through different types of information acquisition methods. Until the evolution of motion sensing devices, the research on HAR [1-3] has been carried with traditional RGB videos which are generally acquired through normal cameras. However, recognizing human actions in RGB images is a challenging issue due to several issues like illumination variations, complex backgrounds, and clothing color, which makes the segmentation of human body much difficult in every scene or picture. Particularly, the RGB images won't have depth clues about the motion of body which has a significant effect on the action recognition.

Recently, the emergence of depth sensors such as Microsoft Kinect sensors has gained much popularity in tracking and capturing the motion in real time applications. After several years of development, research on HAR has made a series of important progress, among which one aspect is most prominent, i.e., the change of the type of information used, from the traditional RGB to the current and popular RGB-D [4-6]. The Kinect sensors have obtained a widespread applicability in so many commercial games because it is inexpensive and also can extract the full-body motions from a general user. The main advantage of depth sensor is its ability to capture the depth and color information simultaneously. The depth cameras ensure depth data as well as color images sequence in real time, which makes the action recognition system more realistic and solves the traditional problems with RGB videos. The RGB-D videos are more advantageous than the traditional RGB videos in several aspects like; 1) The depth cues in RGB-D videos are insensitive to illuminations variations and they can capture the videos even in dark environments also, 2) Depth videos can provide depth data while the traditional ones can't and 3) Texture and color variations are not present in the depth videos which make the action unit detection easier. Due to these advantages, the recent HAR studies have directed in an advanced direction and also motivated several researchers to develop different approaches.

C. Contributions

Unlike most existing papers that focus only on single-modality–based literature surveys, this paper provides an extensive review of human action recognition methods utilizing both depth maps and joint postures. Here are the major contributions:

- Provides an extensive literature survey on human action recognition (HAR) methods using both depth maps and joint postures, unlike most existing works that focus on a single modality.
- Reviews approaches based on handcrafted feature extraction as well as deep learning–based feature extraction, offering a comprehensive methodological coverage.
- Presents a summarized comparative analysis of existing HAR methods, highlighting their strengths, limitations, and practical applicability.

D. Organization of paper

The organization of the paper is as follows: Section II presents a detailed literature survey of existing human action recognition methods, covering both depth map–based and joint posture–based approaches. Section III provides a summarized discussion of the surveyed works, highlighting their respective advantages and limitations to offer a balanced evaluation of the current state of the field.

II. LITERATURE SURVEY

Due to the widespread applicability and huge significance of HAR in different applications, the research on HAR has been started in the 80's only. Most of the past HAR researchers considered RGB videos as input [7], [8]. However, recognizing action from RGB data is a challenging task due to several problems like color variations, illumination variations,

and complex backgrounds etc. The HAR system can't discriminate between action unit and background unit if both are under motion. In such scenario, the segmentation of only action unit is a big challenge in turn it results in poor recognition performance. Moreover, the RGB videos won't provide any cues about the motion. Hence, the current research in HAR has been diverted to the RGB-D data acquired with depth sensors, for example Microsoft Kinect Sensor. RGB-D data provides depth information about the motion and also ensure uniformity in the color information. The RGB-D videos are more advantageous than RGB data in three aspects; they are illumination invariant, motion representation through an additional depth coordinate and absence of texture and color variations [9]. Due to these advantages an extensive research has been carried out within the span of ten years [10-12]. However, due to the availability of two different data modalities in RGB-D such as Depth maps and Skeleton joints, the existing research is partitioned into two parts; they are 1) Depth Maps based HAR and 2) Skeleton Joints based HAR.

2.1 Depth Maps

Under this section, we explore the existing HAR methods those have considered Depth maps as input data modality. Unlike the pixel in RGB videos which was represented only two coordinates, the pixel in depth videos is represented with three coordinates; the additional coordinate explores the depth value of a pixel which is nothing but the distance of the object from a viewpoint. Figure.1 shows an example action video with depth maps in which we can see only two regions namely foreground and background. Whatever the additional regions present in the background, they are all considered as background and represented as a single region.

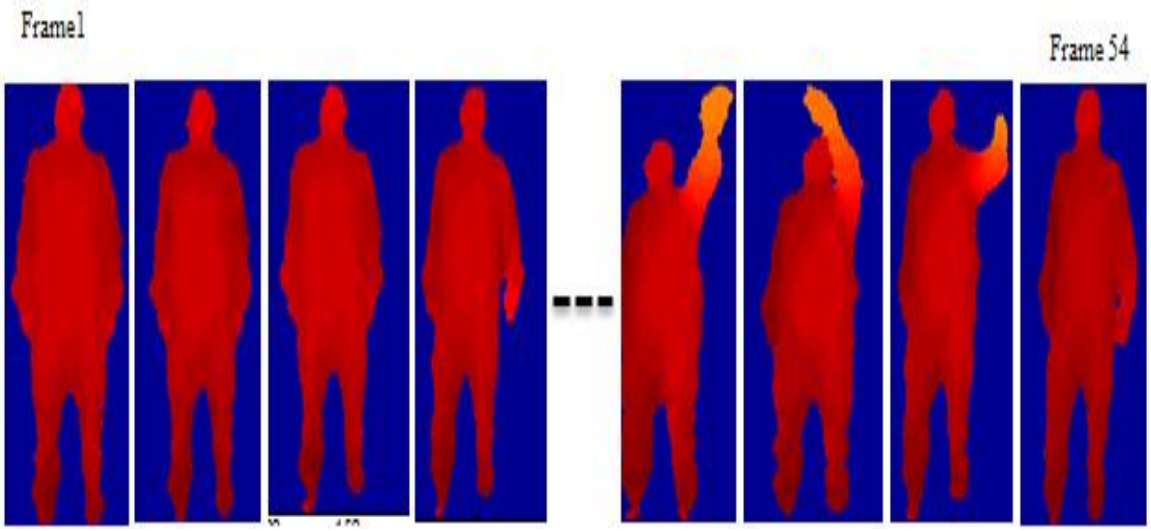


Figure.1 an example action (*right hand wave*) represented with depth maps

The major advantage with depth maps is their robustness to light variations and gives a real depth value. Hence, some of the researchers considered it as an input for HAR system. The entire methodology is accomplished in two phases, they are feature extraction and classification. Based on the features extracted, the existing methods are categorized into two groups; they Handcrafted features and deep learning based feature. The complete details about those existing methods are explored in the following sub-sections;

2.1.1 Handcrafted Features

Under this category, the traditional computer vision and image processing algorithms are applied to extract motion related features from depth action videos. Among the several handcrafted features, Depth Motion Maps (DMM) is found as one of effective motion descriptor initially introduced by Yang *et al.* [13] to recognize human actions. The authors generated DMM by projecting depth maps onto three orthogonal panes (Front, Side and Top) and accumulating the global motion through entire video sequence. Histogram of Oriented Gradient (HOG) features are then derived from DMM descriptor to represent an action and then fed to Linear Support Vector Machine (SVM) for classification.

Similarly, D. Kim *et al.* [14] also computed DMM but for only two planes, they are side plane and front plane and named as "Depth Motion History (DMH)" & "Depth Motion Appearance (DMA)" respectively. Finally, for classification, they also employed SVM algorithm. Unlike these methods, Chen *et al.* [15] generated DMM by accumulating the absolute deviation of successive frames for a depth action sequence. For classification purpose, they employed a L2-regularized classifier with collaborative representation of Distance Weighted Tikhnov Matrix (WTM).

Chen *et al.* [16], [18] enhanced their earlier version of DMM by proposing a segmented DMM computation. Initially, they segmented the video sequence into several overlapping segments and then measured DMM on each segment. They then applied Local Binary Patterns (LBPs) on each DMM to explore the texture information of action. Finally, Fisher Kernel (FK) is applied for encoding the LBP descriptor and then passed to "Kernel-Based Extreme Learning Machine (KELM)" to predict the action. A further enhancement is proposed by Chen *et al.* [17] by proposing a new fusion scenario. They employed fusion at both feature level and decision level. However, the size of segment must be adaptive because each action video has its own length. A common segment size is not an appropriate solution for an effective HAR. R. Azad *et al.* [19] applied Multi-level Temporal Sampling (MTS) and generated three types of sequences with different lengths. Then they calculated Weighted DMM for each temporal sequence to reduce inter-class similarities. Then WDMM is processed for feature extraction through HOG and LBP.

Even though DMM can capture Spatio-temporal depth cues related to motion, it neglects static information. By including static formation, XuWeiyao*et al.* [20] proposed a new Model called as Motion and Static Mapping (MSM) which uses Static History Image (SHI) and Motion History Image (MHI) to describe an action through static and motion postures respectively. Besides MSM, they also proposed a Multi-Frame Select Sampling (MFSS) which captures key frames based on the motion energy. MSM is applied for all the three planes and encoded them with LBP followed by Fisher Kernels. For classification, they employed KELM algorithm. An extended LBP called as Discriminative completed LBP (DiscLBP) is proposed by Wu Li et al. [21] to represent the texture features of an action after the computation of DMM for a give input action video. At classification, they employed two machine learning algorithms namely Collaborative Representation classifier (CRC) and ELM.

Tianjin Yang *et al.* [22] proposed a "Multi-label subspace Learning (MLSL)" mechanism for action recognition from depth maps and named it as "Depth Sequential Information Entropy Maps (DSIEM)". DSIEM represented an action through Spatio-temporal features in which stitching and Entropy were employed to describe temporal and spatial features respectively. After representing the action in a single image, they computed HOG and passed to SVM for action prediction.

Wang *et al.* [23] focused on the provision of noise invariance, and robustness against translational and temporal misalignments and described an action and interaction through a novel features called as Actionlet Ensemble. These features can also ensure robustness against intra-class variations. But, they failed at acquiring the typical motion shapes from the human body. Next, O. Oreife*et al.* [24] developed a new motion descriptor called a Histogram of 4D normals (Hon4d) which describe an action through four features namely two spatial coordinates, depth and time. They computed the Histograms by projecting the initial quantized 4D space distributed on the orientations of several surface normals. Similarly, X. Yang *et al.* [25] proposed a new local motion and shape descriptor by computing the polynormals from the clustered hyper surface normal in a depth action sequence. An adaptive Spatio-temporal pyramid is constructed to acquire the global motion information which sub-divides the depth action video into a set of grids characterized with space-time attributes. Then they proposed a new scheme called as "Super Normal Vector (SNV)" which is a simplified version of Fisher Kernel and aggregates the low level polynormals.

A. W. Vieira *et al.* [26] aimed to handle Intra-class variations in HAR and proposed Spatio-Temporal descriptor called as "Space-Time Occupancy Pattern (STOP)". STOP initially segments the action video into temporal 4D grid and describes each gird with respect to spatial and temporal attributes. A one more 4D based action descriptor called as "Random Occupancy pattern (ROP)" is proposed by J. Wang *et al.* [27] to handle the noise and occlusions. ROP employed a Sparse Feature Encoding method after sampling the extremely larger learning space. Since ROP seeks an additional parameter tuning, it ensured resilience against noise and occlusions in action videos.

By considering the success of "Space-Time Interest points (STIPs)" [29] in the traditional RGB based HAR, L. Xia and J. Agarwal [48] proposed an extended version called as Depth STIPs (DSTIPs) which an effective for the suppression of noises. An additional feature called "Depth Cuboid Similarity Feature (DCSF)" is employed to describe the depth of local 3D cuboid surrounded by DSTIPs. Each action is represented with a set of Bag-of-Words (BoW) and a codebook is constructed with the help of Euclidean Distance assisted K-mean clustering algorithm. Then the cluster centers are used to define the Spatio-temporal code word.

2.1.2 Deep Learning Features

With the advent of advanced hardware equipment and spurious growth of GPUs, Deep Neural Networks (DNNs) attained an increased attention in the computer vision research. Under deep learning category, for HAR, most of the researchers employed 2D CNNs [50]. The main reason behind the vast utilization of CNNs is their ability to make the system rich in learning informative features which is not possible with traditional handcrafted methods. Due to this reason, many researchers extended the CNNs for HAR [31]. In this subsection, we review some of deep learning based HAR methods briefly.

J. Chen *et al.* [32] adapted domain and utilized the features learned from RGB videos for the depth videos. They applied a standard CNN model to perform feature extraction followed by action recognition. Similarly, J. Imran and P. Kumar [53] proposed a 4 channel Deep CNN model in which the first channel classifies the action based on MHI and the remaining three channels classifies based on 3D point clouds calculated from DMM on three orthogonal planes.

P. Wang *et al.* [34] derived three compact representations from depth maps for action recognition; they are namely "Dynamic Depth Motion Normal Images (DDMNI)", "Dynamic Depth Normal Images (DDNI)" and "Dynamic Depth Images (DDI)". These dynamic images are generated from segmented sequence of depth maps through Hierarchical Bi-Directional Rank Pooling (HBRP). Over the obtained representations, they applied a ConvNets for action prediction.

Y. Hou*et al.* [35] Proposed "Spatially and Temporally Structured Dynamic Depth Images (STSDDI)" to aggregate coarse to fine level motion information from depth sequence. STSDDI applied HBRP model [36] to represent depth video in three pairs of Structure Dynamic Images (DIs) at body, part and joint levels. Each level is fed as input to one ConvNet model which is different from earlier works those used only one ConvNet model.

To preserve the temporal information in DMM, Wang P *et al.* [37] proposed A "Hierarchical DMM (HDMM)" with three channel deep CNNs. Initially, the depth map is rotated and then a weight is assigned for every temporal scale DMM and then fed to three channel ConvNets. ImageNet called deep learning model is employed. Next, P. Wang *et al.* [38] extended HDMM by proposing a Weighted HDMM (WHDMM) with three channel ConvNets. After the computation of WHDMM from depth sequence, they applied a pseudo color images and then fed to ConvNets individually.

M. Al-Faris*et al.* [39] introduced fuzzy concept into the DMM and derived a new version called as "Fuzzy Weighted Multi-Resolution DMMs (FWMDMMs)" to represent an action through motion information in segmentation view. At first, the input depth action video is segmented into multiple streams based on time scales and then each stream is processed for DMM computation. Further, the DMMs are assigned fuzzy weights in three ordinations namely Central, Reverse and Linear to signify the motion attributes. At feature extraction and classification, they employed a deep CNN model.

Li Jiang *et al.* [40] employed a Multi-view CNN model which consists of three CNN models used to work for three views such as front, side and top. At the classification, the fully connected layers output is merged as a softmax regression to predict the action. Extending the LBP to Local Ternary Pattern (LTP), Li Z. *et al*. [41] proposed an action descriptor through DMM. Over the DMM, they applied LTP instead of LBP to encode the texture features. Finally for classification, they approached to a deep CNN model.

Unlike the above method those encoded the depth action sequence into a single image, Hanbo Wu *et al*. [42] proposed a Dynamic Image Sequence (DIS) which focused on Spatio-Temporal Attention Points and describes the action through local Spatio-temporal dynamics. Then they modeled Channel Attention (CA) model based CNNs for feature extraction followed by classification.

2.2 Skeleton Joints or Body Postures

Under this section, we explore the existing HAR methods those have considered Skeleton Joints as input data modality. The skeleton joints represents an action data in more compact way because it makes to disappearance the background noises and illumination variations. Next, the skeleton is represented inherently with 3D joint positions and consumes very less memory. Alongside, with the invention of powerful depth sensor like Microsoft Kinect sensor, accessing the skeleton joints position is very much easy and flexible. With the motivation of capturing the skeleton data through Kinect sensors in real time, so many works [43-45] are developed for action analysis through skeleton videos. Figure.2 shows an example action with a set of frames where each frame with represented with 3D skeleton joints. Here each frame is represented with 20 joints and each joint is represented with three positions like J = $J = (x, y, z)$. From the Figure, it can be see that there are only joints and the remaining data is completely discarded. Such kind of representation costumes very less memory to store an action even with larger time periods. Microsoft Kinect Sensor is used to acquire skeleton joints by attaching the markers to the human body at different joints, as shown in Figure.3.

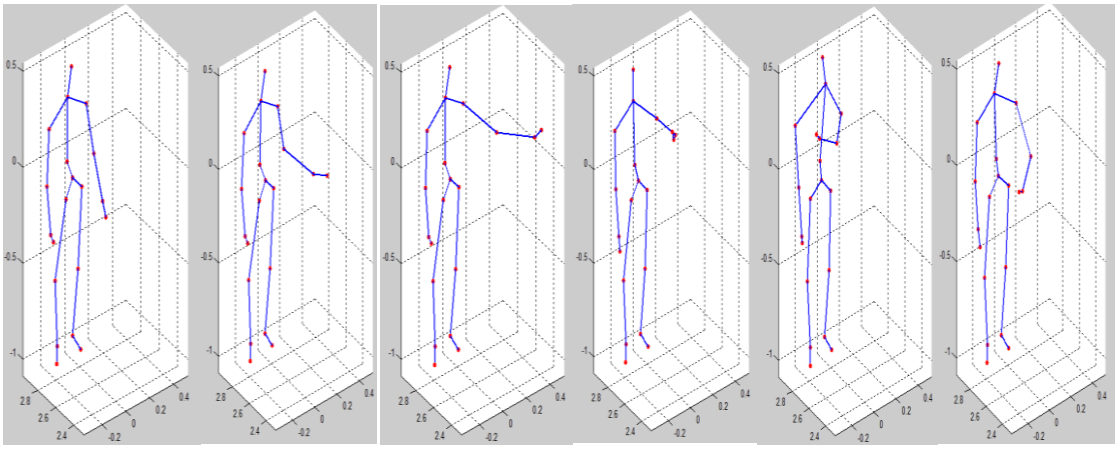
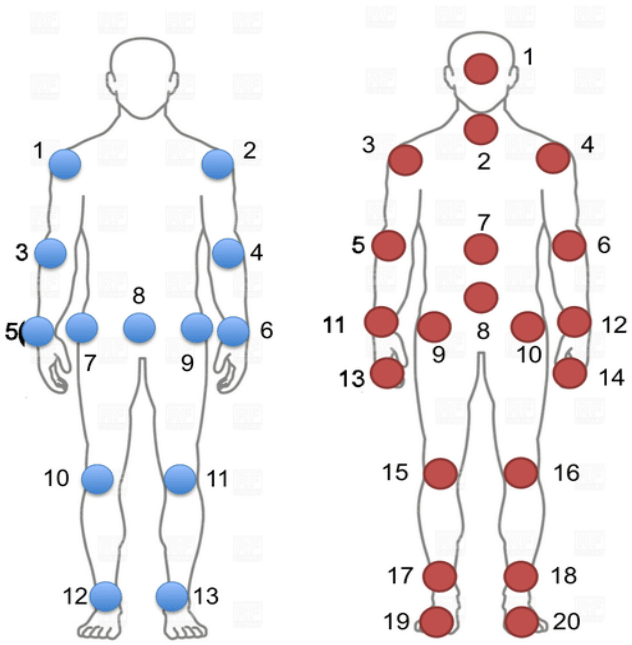


Figure.2 Action: *Horizontal arm Wave*, representing with 3D skeleton joints



(a) (b)

Figure.3 Different Skeleton joint configurations (a) 13 joints and (b) 20 Joints

Due to several advantages with skeleton joints, most of the current research on HAR has been used them as an input data. Towards this research, so many researchers have worked and proposed several representation methods. Some authors employed handcrafted methodswhile some authors employed deep learning methods. In the former category, the authors focused on the development of new action descriptor and they used mostly standard deep learning pre-trained models for classification. In the second category, the deep learning models are employed for both feature extraction as well as for classification. In this section, we brief out different methods proposed in the past.

2.2.1. Handcrafted Methods

Xia. L. *et al.* [46] suggested a compact action descriptor based on skeleton joints and it is named as "Histogram of 3D joint Locations (HOJ3D)". At first, they converted the Cartesian coordinates of skeleton joints into spherical coordinates. Next, they applied "Linear Discriminant Analysis (LDA)" for re-projection and then applied clustering mechanism to cluster them into different posture visual words. Discrete Hidden Markov Model (DHMM) [68] is applied on visual words to assess the temporal evolutions. This method shows a better recognition performance even for the action acquired under multiple views. However, the accomplishment of DHMM does not have any physical meaning. Due to this reason, HOJ3D can't provide a qualitative insight that correlates the action with the original meaning of an action. To address these problems, F. Ofli et al. [47] proposed an informative joint descriptor called as "Sequence of Most Informative Joints (SMIJ)". Before obtaining SMIJ, the input action sequence is subjected to clustering and obtains different clusters with different poses. Then

they are subjected to histograms computation. At every time instant, they select few skeletal joints automatically that are presumed to be more informative. The selection is done based on several statistical measures such as joint's maximum angular velocity, variance and mean of joint angles etc.

X. Yang et al. [49] developed a skeleton joint descriptor for HAR based on the difference between the positions of skeleton joints. They computed Eigen Joints that combines overall dynamics, motion property, and static posture information. Further, to remove noise frames and to lessen the computation cost, they proposed an "Accumulated Motion Energy (AME)" metric. "Naïve-Bayes-Nearest-Neighbor (NBNN)" is used to perform the classification.

Vemulapalli et al. [50] modeled an action descriptor from skeleton joints based on their 3D geometric relationships through different rotations and translations in 3D space. This kind of representation represents the actions as curves in the lie group. However, the classification is tough for curve classification. Hence, the curve is transformed into algebra and then fed to a combinational classifier. They employed totally three classifiers namely "Dynamic Time Warping (DTW)", "Fourier Temporal Pyramid (FTP)", and "Support Vector Machine (SVM)" algorithms.

M. E. Hussein et al. [51] described the skeleton action frame through location assisted covariance matrix. Further, for temporal exploration, they employed covariance between time segmented frames. They considered hierarchical fashion and the descriptors are of fixed length irrespective of the action sequence length. M. Jiang et al. [52] contributed towards the elimination of unrelated joints and construct informative joint set for each action. To make the recognition system robust, they transformed the skeleton to a standard skeleton by applying normalization through scaling, rotation and translation. The skeleton similarities are assessed through the skeleton context which is a multi-scale binned pairwise spatial distribution of informative joints. Next, they quantized through Affinity propagation (AP) [57] for clustering the feature vectors and then fed to Conditional Random Fields (CRFs) for classification. However, the methods proposed in [46] and [52] assumed that the original skeletons are perpendicular to ground plane. In general, these methods transformed each skeleton into a specific coordinate system. But they are much sensitive to noises in skeleton and results in loss of Spatio-temporal relationship between frames of an action sequence.

H. Rahmani*et al.* [53] aimed to achieve robustness for different scales, noises, speeds and view points and proposed an action descriptor called as "Histogram of oriented principal components (HOPC)". HOPC is computed at every 3D point by the projection of three scaled Eigen vectors within its Spatio-temporal volume called as dodecahedron. HOPC determines the Spatio-temporal key points (STK) both in local and global fashions so that the system can achieve view invariance. Chen *et al.* [74] also concentrated on the speed invariance and intra-class variance and developed a 2-level hierarchical descriptor for 3D skeleton action sequences. At first, they segmented the entire frame into five segments to mitigate the intra-class variance. Next, every cluster is subjected to the computation of motion attributes and they are fed to Graph assisted action recognition through "Maximum Likelihood Estimation (MLE)".

E. Cippitelli [55] proposed a posture vector computation for each skeleton frame based on the spatial features in the 3D skeleton coordinates. A normalized distance is computed between the torso, neck and remaining joints such that it becomes speed and motion independent. Next, the feature selection process is employed through K-means clustering algorithm followed by the determination of final feature vector. Multi-Class SVM is employed for classifying the actions. D. Warchol and K. Tomasz [56] proposed an action descriptor that encodes the angular correlation between the bone pairs. Then they are integrated with distance features used to signify the relationship between skeleton joints. Five classifiers are employed for classification; they are namely (1) "LogDet Divergence Based Metric Learning with Triplet Constraints (LDMLT)", (2) "Bidirectional Long Short-Term Memory Network (BiLSTM)", (3) "Fully Convolutional Network (FCN)", (4) "DTW with City Block Distance (DTW-CBD)", and (5) "DTW with Euclidean Distance (DTW-ED)".

One more way to provide view invariance is through Self-similarity Matrix (SSM) [58] computation between frames of an action sequence. Y. Hsu *et al.* [59] developed a view invariant action recognition model by exploring the advantages of SSM. SSM computes the similarities between frames through Euclidean distance and the obtained distances are formulated into a 2D symmetric matrix called as Spatio-Temporal matrix (STM). For the recognition of action, they employed SVM after describing the action with STM using the pyramid-structural bag-of-words (BoW-pyramid). With the availability of exact positions of skeleton joints, the skeletons can be made strictly view invariant after the transformation.

2.2.2. Deep Learning Methods

In this category, several deep learning methods are developed those can perform both feature extraction and classification. Among the several deep learning models, CNNs have gained better recognition in recognizing human actions. Y. Du *et al.* [60] developed a top-to-bottom hierarchical structure for the recognition of human actions based on skeleton joints. At first, they construct a matrix by associating the location coordinates of each skeleton joint at each instant. Then the vectors of the matrix are arranged in a chronological order to solve the variable length problem and finally fed to CNN model that extracts features and also recognizes the action.

P. Wang *et al.* [61] utilized the ConvNets [82] in their method for action recognition based on skeletons. They introduced a new descriptor named as "Joint Trajectory map (JTM)" that encodes the Spatio-temporal relationship between skeleton frames. Some authors applied recurrent neural networks (RNN) for action recognition in which the "Long-Short

Tem Memory (LSTM)" is used to describe the temporal related evolutions in skeleton action sequences.

X. Diao*et al.* [63] suggested a new RNN model named as "Multi-Term Attention Networks (MTANs)" which can extract the temporal features at different scales. This network consists of MTA-RNN and ST-CNN. Mengyuan Liu *et al.* [84] proposed a sequence based view invariant transform to cope up with view point variations in HAR. The view invariance is achieved through only torso joints which are limited in number and they have serious restrictions on the rotations. Hence, the recognition system is sensitive to motions with similar movements even at same angles.

J. Liu *et al.* [65] proposed a new version of LSTM network called as Global Context Aware Attention LSTM (GCA-LSTM) for skeleton based action recognition. They considered the Global Memory Cell to select the informative joints from each skeleton frame. Further, they also introduced a recurrent attention mechanism to enhance the capability of their network and the training was done in a step-by-step process. A similar method is proposed by J. Liu *et al.* [66] by introducing a Gating Mechanism with LSTM to deal with noisy skeleton. However, the noise due to similar movements makes the recognition system more confused and results in larger number of false positives for large scale datasets.

QiangNie*et al.* [67] proposed two motion features namely Euclidean distance matrix between joints (JEDM) and Joint Euler Angles (JEAs) to cope up with View invariant, noise and occlusions in HAR with skeleton videos. However, they restricted the motion limits to certain angles which is not a vital solution because the angle of rotation is totally independent and it varies based on several factors like bone length, muscle strength, age of person, gender etc.

Some researchers applied the Graphical Convolutional Networks (GCNs) to represent an action through skeleton joints. For example, Yan *et al.* [68] developed a new GCN model called as Spatio-Temporal GCN (ST-GCN) to describe an action through its Spatio-Temporal features. However, the GCN seeks the manual setting network topology, in addition, the number of layers, and input samples are static. Moreover, the ST-GCN can't investigate the special attributes like bone length, and bone movement directions. Further, Shi *et al.* [69] introduced an Adaptive two-stream GCN which trains the HAR system uniformly from one end to other end. Zhang *et al.* [90] encouraged to use the edges of the graph which maps with the bones of human skeleton. An edge is described with the help of Spatio-Temporal neighbour edges which can explore the relation between bones and consistency of movements in the action video. Further, a Graph node CNN and Graph Edge CNN are constructed based on the intermediate layers shared [71].

In general, the deep learning models can extract the required set of features automatically which denotes the local patterns of image, thus it can encode the spatial and temporal relationship of an action sequence. However, for skeleton joints, the input data size is very limited and hence the CNN models create a lot of ambiguity between features of different actions. So, there is a need of an effective joint descriptor to describe a skeleton sequence which can make the recognition system robust to viewpoints, speed, movements, etc.

2.3 Hybrid Methods

Even through the individual model based methods has gained significant accuracy; they can't solve the problems with other models. Some of the problems with individual methods are listed as follows;

1. The existing skeleton based HAR methods mostly concentrated on the single view, i.e., the actions used for training and testing is acquired from only one view. In such case, the same action captured in different views may or may not get recognized. HAR in such instances is called as Cross view HAR which is tough task. Moreover, in skeleton based methods, the past researchers didn't concentrate on the redundancy of joints which constitutes computational burden on the recognition system.
2. Depth images are composed of different types of noises like small body shaking movements, jumbled objects, cluttered backgrounds, ghost shadows etc. Due to these noises, the action descriptor comprises with fake moving pixels which consequences to less recognition accuracy.

Hence, some of the recent researchers considered multiple data models for HAR. The hybrid methods follows different strategic fusion processes to produce the final action label. Broadly they are two types of fusion processes namely Early fusion and late fusion. In the former type, the HAR system fuses the action descriptors and represents an action with combined descriptor. In the case of late fusion, the HAR system is individually tested and the obtained results are fusion through different fusion rules. Broadly there are three types of fusion rules namely maximum fusion, average fusion, and product fusion. Different methods followed different types of fusion strategies and produced action prediction results. Some of the Hybrid methods are explored here;

A. Kamel*et al*. [72] considered body postures and depth maps as input data modalities and proposed a new CNN model with three channels. Further, they proposed two new action descriptors such as "Depth Motion Image (DMI)" for depth maps and "Motion Joint Descriptor (MJD)" for body postures. They trained the system with three CNN channels and the obtained results are fused to get the action prediction value. For final action prediction, they suggested two fusion rules namely max fusion, and product fusion. DMJ is a basic motion descriptor which doesn't have any additional capabilities to label the non-motion regions. Due to this reason, it has less performance at Cross subjects.

Sun Yat-sen [73] focus on exploring modality-temporal mutual information for RGB-D action recognition. In order to learn time varying information and multi-modal features jointly, they propose a novel deep bilinear learning framework. In

the framework, bilinear blocks are proposed which consist of two linear pooling layers for pooling the input cube features from both modality and temporal directions, separately. To capture rich modality-temporal information and facilitate our deep bilinear learning, a new action feature called modality-temporal cube is presented in a tensor structure for characterizing RGB-D actions from a comprehensive perspective.

Y. Fan *et al.* [74] proposed a cross attention module for action recognition which is an integration of cross attention and self-attention branches. This approach can extract informative joints which are highly correlated with the context of scenario. Q. Cheng *et al.* [75] considered RGB and Depth data modalities and proposed a "Spatio-temporal information aggregation module (SITAM)" which utilizes CNNs to acquire Spatio-temporal information from input data. Further, they introduced a "Cross Modality Interactive module (CMIM)" to aggregate the multi-modal complementary information completely. Finally, an integrated model called as "Multi-modal interactive network (MIMINet)" by fusing the SITAM and CMIM.

Table.1 provides a summarized comparison between existing HAR methods. Compared to the single data modality, the multiple data modalities ensure an increased recognition performance in HAR. However, there are several constraints; 1) inappropriate data modalities creates additional storage burden. 2) Pure deep learning algorithms can't ensure the perfect discrimination between actions. 3) Fake moving pixels are not nullified. On the other side, in the case of depth maps, the past methods didn't focus on the nullification of fake moving pixels.

Table.1 Summarized Comparison between existing HAR methods

Author	**Paper Title**	**Publication & Year**	**Methodology**	**Limitations**
J. Liu et al. [66]	Skeleton based action recognition using Spatio-temporal LSTM network with trust gates.	*IEEE Trans. Pattern Anal. Mach. Intell.*, 2018.	1. ST-LSTM for action recognition. 2. Trust gate to minimize the positional errors of locations.	No knowledge provision for system about same action under different views
A. Shahroud, J. Liu, T.T. Ng, G. Wang [76]	NTU RGB+D: a large scale dataset for 3D human activity analysis.	*Computer Vision and Pattern Recognition (CVPR),* 2016.	1. Body partitioning 2. Part aware LSTM in Recurrent Neural Networks (RNNs) for classification	View invariance is not provided
R. Vemulapalli, F. Arrate , R. Chellappa [46]	Human action recognition by representing 3D skeletons as points in a lie group	*CVPR*, 2014	1. 3D geometric relationship through Euclidean distance 2. Curve Formulation 3. Support Vector Machine for Classification	Highly sensitive to inter-class similarities
L. Xia, C. C. Chen, and J. Aggarwal [50]	View invariant human action recognition using histograms of 3D joints,	*CVPR*, 2012	1. Histograms computation over skeleton Joints 2. Linear Discriminant Analysis (LDA) for discrimination provision 3. SVM for classification	No knowledge provision for system about same action under different views
XiaoleiDiao et al. [63]	Multi-Term Attention Networks for Skeleton-Based Action Recognition	*Appl. Sci.*, Vol.10, no.5326, pp.1-19, 2020	1. MTA-RNN trained individually by three coordinates. 2. ST-CNN is applied for Spatio-temporal features extraction	Larger false positive rate at different views
Mahmoud Al-Faris et al . [39]	Deep Learning of Fuzzy Weighted Multi-Resolution Depth Motion Maps with	*Journal of Imaging*, Vol. 5, issue. 82, 2019	1. Fuzzy weighted multi-resolution Depth Motion Maps (FWMDMMs). 2.Convolutional Neural	Small body shaking movements are also encoded as motion features.

	Spatial Feature Fusion for Action Recognition		Networks	Weight increased with time is not an optimal solution
Tianjin Yang et al. [22]	Depth Sequential Information Entropy Maps and Multi-Label Subspace earning for Human Action Recognition	*IEEE Access*, Volume 8, 2020, pp. 135118- 135130	1. Depth Sequential Information Entropy Maps (DSIEM) 2. Histogram of Oriented Gradients (HoGs) 3. Support Vector Machine (SVM)	Entropy can't suppress the sides effects like noises, movements due to clothes etc.
A. Kamel*et al*. [72]	Deep Convolutional Neural Networks for Human Action Recognition Using Depth Maps and Postures	*IEEE Trans. on Systems, Man, and Cybernetics: Systems*, 49(9), 1806-1819, 2019	1. Depth Motion Image (DMI) for depth maps and 2. Motion Joint Descriptor (MJD) for body postures	DMI cannot suppress the noises and MJD is not resilient against inter-class variations
Q. Cheng *et al.* [75]	Spatial-Temporal Information Aggregation and Cross-Modality Interactive Learning for RGB-D-Based Human Action Recognition	*IEEE Access*, Vol. 10, pp.104190-104201, 2022	CMIM Provides Spatio-temporal interactive information between RGB and Depth data modalities.	View invariance is not addressed

III. Discussion and Conclusion

The reviewed literature on human action recognition presents a variety of skeleton-based, depth-based, and multimodal approaches, each with distinct methodologies and limitations. Liu et al. (2018) employed a Spatio-Temporal LSTM (ST-LSTM) with trust gates to reduce positional errors, though the system lacked knowledge provision for recognizing the same action from different viewpoints. Shahroud et al. (2016) introduced the NTU RGB+D dataset and used body partitioning with part-aware LSTM in RNNs, but without addressing view invariance. Vemulapalli et al. (2014) modeled 3D skeletons as points in a Lie group using geometric relationships, curve formulation, and SVM classification, yet their method was highly sensitive to inter-class similarities. Xia et al. (2012) computed histograms over skeleton joints, applied Linear Discriminant Analysis (LDA), and classified using SVM, but similarly lacked cross-view knowledge provision. Diao et al. (2020) proposed Multi-Term Attention Networks (MTA-RNN) trained on three coordinate sets with ST-CNN for spatio-temporal feature extraction, though it suffered from higher false positives at different views. Al-Faris et al. (2019) utilized fuzzy weighted multi-resolution depth motion maps with CNN-based spatial feature fusion, but small body shakes were wrongly encoded as motion, and increasing weights over time proved suboptimal. Yang et al. (2020) developed Depth Sequential Information Entropy Maps (DSIEM) combined with HoG features and SVM classification, which could not effectively suppress noise and clothing-induced movements. Kamel et al. (2019) integrated Depth Motion Images (DMI) for depth maps with Motion Joint Descriptors (MJD) for postures, yet DMI was noise-prone and MJD lacked robustness against inter-class variations. Finally, Cheng et al. (2022) proposed a cross-modality interactive learning framework (CMIM) for RGB-D action recognition, enabling spatio-temporal feature sharing between RGB and depth modalities, though without addressing view invariance.The reviewed human action recognition methods face several persistent challenges that limit their robustness and generalization in real-world scenarios.

1.Viewpoint dependency - is a major issue — many approaches (e.g., Liu et al., Xia et al., Cheng et al.) cannot reliably recognize the same action when captured from different camera angles, making them less effective in unconstrained environments.

2. Inter-class similarity - remains problematic — methods such as Vemulapalli et al. and Kamel et al. struggle when different actions have similar motion patterns, leading to misclassification.

3. noise sensitivity - is a recurring challenge — depth-based and skeleton-based techniques (e.g., Yang et al., Kamel et al.) often misinterpret irrelevant variations such as clothing movement, sensor noise, or small unintentional body shakes as meaningful action cues.

4. False positives and overfitting to motion cues - occur when minor or unintended motions (Al-Faris et al.) are encoded as significant features, reducing system precision.

5. Temporal weighting and feature representation limitations - can lead to suboptimal learning — for instance, increasing motion weights over time (Al-Faris et al.) or relying on entropy measures that can't filter out irrelevant variations (Yang et al.) results in noisy feature spaces.

6. Dataset and modality constraints - affect generalization — while some models use large datasets like NTU RGB+D, others are trained on limited scenarios, and many methods do not leverage cross-view or cross-modality robustness effectively.

In summary, these methods struggle with view invariance, inter-class discrimination, noise resilience, temporal feature reliability, and domain generalization, all of which are critical for deploying action recognition systems in practical, uncontrolled settings.

Future Scope: The future of human action recognition using depth maps and skeleton postures is to get enhanced accurate, robust real time systems expending the potential of HAR in various field using advanced deep learning focus will be on Cross -Modal fusion techniques with different data representation can provide a comprehensive understanding of human actions.

References

1. Hong-Bo Zhang, Yi-Xiang Zhang, BinengZhong, Qing Lei, Lijie Yang, Ji-Xiang Du, and Duan-Sheng Chen, "A Comprehensive Survey of Vision-Based Human Action Recognition Methods", Sensors 2019, 19, 1005.

[2]. Yu, K.; Yun, F. Human Action Recognition and Prediction: A Survey. arXiv2018, arXiv:1806.11230

3. Dawn, D.D.; Shaikh, S.H. A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector.*Vis. Comput*. 2016, 32, 289–306.

4. B. Liang and L. Zheng, "A Survey on Human Action Recognition Using Depth Sensors," *2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Adelaide, SA, Australia, 2015, pp. 1-8,

5. Chen, C., Jafari, R. &Kehtarnavaz, N. A survey of depth and inertial sensor fusion for human action recognition. *Multimed Tools Appl* 76, 4405–4425 (2017)

6. Morshed, MdGolam, Tangina Sultana, AftabAlam, and Young-Koo Lee. 2023. "Human Action Recognition: A Taxonomy-Based Survey, Updates, and Opportunities" *Sensors* 23, no. 4: 2182.

7. Beddidar, D. R., Nini, B., Sabokrou, M. et al. "Vision-based human activity recognition: a survey", Multmedia Tools Appl., 79, 30509-30555, 2020.

8 Shugang Zhang, Zhinqiang Wei, JieNie, Lei Huang, Shuang Wang, Zhen Li, "A review on human activity recognition using vision-based method", Journal of Healthcare Engineering, Vol.2017, Article ID 3090343, 31 pages, 2017.

9. L. Chen, H. Wei, and J. Ferryman, "A survey of human motion analysis using depth imagery," *Pattern Recognition Letters*, 2013, pp. 1995-2006.

10. P. Wang, W. Li, P. Ogunbona, "RGB-D-based human motion recognition with deep learning: a survey". *Computer Vision & Image Understanding*, 171: 118–139, 2018.

11. Shaikh MB, and Chai D. "RGB-D Data-Based Action Recognition: A Review", ***Sensors***. 2021; 21(12):4246.

12. PushpajitKhaire, Praveen Kumar, "Deep learning and RGB-D based human action, human–human and human–object interaction recognition: A survey", *Journal of Visual Communication and Image Representation,* Volume 86, 2022, 103531.

13. Yang, X., Zhang, C., &Tian, Y. (2012). "Recognizing actions using depth motion map- s-based histograms of oriented gradients", In *Proceedings of the 20th ACM international conference on multimedia,* New York, NA, USA, pp. 1057–1060, October 2012.

14. 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), Rome, Italy, pp. 2308–4278, 2014.

15. C. Chen, K. Liu, and N. Kehtarnavaz, "Real-time human action recognition based on depth motion maps".*Journal of Real-time Image Processing*, Vol.12, no. 1, pp.155–163, 2016.

16. C. Chen, M. Liu, B. Zhang, J. Han, J. Jiang, and H. Liu, "3D action recognition using multi-temporal depth motion maps

and fisher vector", in *Proc. of Twenty-Fifth International Joint Conference on Artificial Intelligence*, New York, NA, USA, pp. 3331–3337, 2016.

17. Chen, C., Jafari, R., and Kehtarnavaz, N., "Action recognition from depth sequences using depth motion maps-based local binary patterns", In *Proc., ofIEEE winter conference on applications of computer vision (WACV)*, Waikoloa Beach, pp. 1092–1099, 2015.

18. C. Chen, M. Liu, H. Liu, B. Zhang, J. Han, and N. Kehtarnavaz, "Multi-temporal depth motion maps-based local binary patterns for 3-D human action recognition,'" *IEEE Access*, vol. 5, pp. 22590-22604, 2017.

19. R. Azad, M. Asadi-Aghbolaghi, S. Kasaei and S. Escalera, "Dynamic 3D Hand Gesture Recognition by Learning Weighted Depth Motion Maps," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 6, pp. 1729-1740, June 2019.

20. XuWeiyao, Wu Muqing, Zhao Min, Liu Yifeng2, Lv Bo, And Xia Ting, "Human Action Recognition Using Multilevel Depth Motion Maps", *IEEE Access*, Vol. 7, 2019, pp.41811-41822.

21. 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.

22. Tianjin Yang , ZhenjieHou, Jiuzhen Liang ,YuwanGu, and Xin Chao, "DepthSequential Information Entropy Maps and Multi-Label Subspace Learning for Human Action Recognition", *IEEE Access*, Vol.8, 2020, pp.135118-135130.

23. J. Wang, Z. Liu, and Y. Wu, "Mining action let ensemble for action recognition with depth cameras", *In Proc., of IEEE conference on Computer vision and pattern recognition (CVPR),* Providence, Rhode Island, USA, pp.1290–1297, 2012.

24. Oreifej O, and Liu Z. "Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 716– 723.

25. Yang X, and Tian Y, "Supernormal vector for activity recognition using depth sequences*," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 804–811.

26. Vieira A, Nascimento E, Oliveira G, Liu Z, and Campos M, "Stop Space-time occupancy patterns for 3D action recognition from depth map sequences," *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pp. 252–259, 2012.

27. Wang J, Liu Z, Chorowski J, Chen Z, and Wu Y, "Robust 3d action recognition with random occupancy patterns," *in Computer vision–ECCV*, Springer, pp. 872–885, 2012.

28. L. Xia and J. Aggarwal, "Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera".*In: Proc., of the IEEE conference on computer vision and pattern recognition*, Portland, OR, USA, , pp. 2834–2841, 2013.

29. P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse Spatio-temporal features. In PETS, pages 65–72, 2005

30. LeCun, Y., Bottou, L., Bengio, Y,,&Haffner, P., "Gradient-based learning applied to document recognition. In *Proc. of the IEEE*, Vol.86, No. 11, pp.2278–2324, 1998.

31 Asadi-Aghbolaghi, M., Clapes, A., Bellantonio, M., Escalante, H. J., Ponce-Lopez, V., Baro, X., Escalera, S., "Deep learning for action and gesture recognition in image sequences: A survey". In *Gesture recognition, Cham: Springer*, pp. 539–578, 2017

32. J. Chen *et al.*, "Action recognition in depth video from RGB perspective: a knowledge transfer manner," in *Proc. Conf. Pattern Recognit. Comput.Vision*, Guangzhou, China, pp. 1060911–1060916, 2018.

33. J. Imran and P. Kumar, "Human action recognition using RGBD sensor and deep convolutional neural networks," in *Proc. Int. Conf. Adv. Comput., Commun.Informat.*, Jaipur, India, 2016, pp. 144–148.

34. P. Wang, W. Li, Z. Gao, C. Tang, and P. O. Ogunbona, "Depth pooling based Large-Scale 3-D action recognition with convolutional neural networks," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1051–1061, May 2018.

35. Y. Hou, S. Wang, P. Wang, Z. Gao, and W. Li, "Spatially and temporally structured global to local aggregation of dynamic depth information for action recognition," *IEEE Access*, vol. 6, pp. 2206–2219, 2017.

36. P. Wang, S. Wang, Z. Gao, Y. Hou, and W. Li, "Structured images for RGB-D action recognition," in *Proc. IEEE Int. Conf. Comput. Vision*, Venice, Italy, pp. 1005–1014, 2017.

37 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.

38. Wang, P., Li, W., Gao, Z., Zhang, J., Tang, C. &Ogunbona, P. O., "Action Recognition from Depth Maps Using Deep Convolutional Neural Networks. *IEEE Transactions on Human-Machine Systems*, Vol. 46, No. 4, 498-509, 2016.

39 Mahmoud Al-Faris, John Chiverton, Yanyan Yang, 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.

40 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.

41. Li, Z., Zheng, Z., Lin, F. et al. "Action Recognition from depth sequence using depth motion maps based local ternary patterns", *Multimedia Tools Appl.*, 78, 19587-19601, 2019.

42. Hanbo Wu , XinMa, and Yibin Li, "Convolutional Networks With Channel and STIPs Attention Model for Action Recognition in Videos", *IEEE Transactions on Multimedia*, Vol. 22, No. 9, pp.2293-2306, September 2020.

43. X. Yang, Y. Tian, "Eigen joints-based action recognition using Naive–Bayes-nearest-neighbor", in *Proc. of the Conf. on Computer Vision and Pattern Recognition Workshops*, Providence, RI, USA, 2012, pp. 14–19.

44. Liu, A.A., Nie, W. Z., Su, Y. T., Ma, L., Hao, T., Yang, Z. X., "Coupled hidden conditional random fields for RGB-D human action recognition," *Signal Process.*, vol.112, pp.74–82, 2015.

45. M. Ding, and G. Fan, "Multilayer joint gait-pose manifolds for human gait motion modeling," *IEEE Trans. Cybern.*, vol.45, no.11, pp.2413–2424, 2015.

46. L. Xia, C. C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Proc. of the Conf. on Computer Vision and Pattern Recognition Workshops*, Providence, RI, USA, 2012, pp. 20–27.

47. 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.

48. F. Lv and R. Nevatia., "Recognition and Segmentation of 3D Human Action Using HMM and Multi-class Adaboost", In *Proc. ECCV*, Graz, Austria, 2006, pp.359-372.

49. X. Yang, Y. Tian, "Effective 3D action recognition using Eigen joints," *J. Vis. Commun. Image Represent.*, vol.25, no.1, pp.2–11, 2014.

50. R. Vemulapalli, F. Arrate , R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in: P*roc. of CVPR*, 2014, pp. 588–595.

51 M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban., "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations", in *Proc.IJCAI*, Bejing, China, 2013, pp. 2466–2472.

52. M. Jiang, J. Kong, G. Bebis, H. Huo, "Informative joints based human action recognition using skeleton contexts", *Signal Process. Image Commun.*, vol.33, pp. 29–40, 2015.

53. H. Rahmani, ArifMahmood, Du Q. Huynh, AjmalMian "HOPC: Histogram of oriented principal components of 3D point clouds for action recognition", in *Proc. of ECCV*, Zurich, Switzerland, 2014, pp. 742-757.

54. H. Chen, G. Wang, J. H. Xue, and L.He, "A novel hierarchical framework for human action recognition," Pattern*Recognit.*, vol. 55, pp. 148-159, Jul. 2016.

55. E. Cippitelli, S. Gasparrini, E. Gambi, and S. Spinsante, "A Human activity recognition system using skeleton data from RGBD sensors", *Computational intelligence and Neuroscience*, vol.2016, Article ID 4351435, pp.1-14, 2016.

56 DawidWarchol and Tomasz Kapuscinski, "Human Action Recognition Using Bone Pair Descriptor and Distance Descriptor", *Symmetry*, vol.12, no.1580, pp.1-12, 2020.

57. B.J. Frey, and D. Dueck, "Clustering by passing messages between data points," *Science*, vol.315, no.5814, pp.972-976, 2007.

58. I.N. Junejo, E. Dexter, I. Laptev, P. Perez, "View-independent action recognition from temporal self-similarities," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol.33, no.1, pp.172–185, 2011.

59. Y. Hsu, C. Liu , T. Chen, and L. Fu, "Online view-invariant human action recognition using RGB-D Spatio-temporal matrix," *Pattern Recognit.*, vol.60, pp.215–226, 2016.

60. Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *Proc.ACPR*, Kuala Lumpur, Malaysia, 2015, pp. 579–583.

61. P. Wang, W. Li, C. Li, and Y. Hou, "Action recognition based on joint trajectory maps with convolutional neural networks," arXiv preprint arXiv:1612.09401, 2016.

62. J. Koushik, "Understanding convolutional neural networks," arXiv preprint arXiv:1605.09081, 2016.

63. XiaoleiDiao, Xiaoqiang Li and Chen Huang, "Multi-Term Attention Networks for Skeleton-Based Action Recognition", *Appl. Sci.*, vol.10, no.5326, pp.1-19, 2020.

[64] Mengyuan Liu, Hong Liu, and Chen Chen, "Enhanced skeleton visualization for view invariant human action

recognition", *Pattern Recognition*, vol.68, pp.346-362, 2017.

65. J. Liu, G. Wang, L.-Y.Duan, K. Abdiyeva, and A. C. Kot, "Skeleton based human action recognition with global context-aware attention LSTM networks," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1586–1599, Apr. 2018.

66. J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton based action recognition using Spatio-temporal LSTM network with trust gates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 3007–3021, Dec. 2018.

67. QiangNie, Jiangliu Wang, Xin Wang, Yunhui Liu, "View-Invariant Human Action Recognition Based on a 3D Bio-Constrained Skeleton Model", *IEEE Transactions on Image Processing*, Vol. 28, No. 8, August 2019, pp.3959-3972.

68. Yan, S., Xiong, Y., Lin, D. (2018). "Spatial temporal graph convolutional networks for skeleton-based action recognition", *In Thirty-Second AAAI Conference on Artificial Intelligen*ce, pp. 7444–7452, 2018.

69. Shi, L., Zhang, Y., Cheng, J., Lu, H., "Two-stream adaptive graph convolutional networks for skeleton-based action recognition", *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recogniti*on, pp. 12026-12035, 2019.

70. Zhang, X., Xu, C., Tian, X., Tao, D., "Graph edge convolutional neural networks for skeleton-based action recognition", *IEEE Transactions on Neural Networks and Learning Syste*ms, 31(8): 3047-3060, 2019.

71. Soo Kim, T., Reiter, A., "Interpretable 3d human action analysis with temporal convolutional networks", *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1-9, 2017.

72. A. Kamel, Bin Sheng, Yang Po, Ping Li, and RuiminShen, "Deep Convolutional Neural Networks for Human Action Recognition Using Depth Maps and Postures", *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, Vol. 49, No.9, pp.1806-1819, 2019.

73. J. F. Hu,W.-S. Zheng, J. Pan, J. Lai, and J. Zhang, ``deep bilinear learning for RGB-D action recognition," In: *Proc. of Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, pp. 335-351, 2018.

74. Y. Fan, S. Weng, Y. Zhang, B. Shi and Y. Zhang, "Context-Aware Cross-Attention for Skeleton-Based Human Action Recognition", *IEEE Access*, Vol. 8, pp. 15280-15290, 2020.

75. Q. Cheng, Z. Liu, Z. Ren, J. Cheng and J. Liu, "Spatial-Temporal Information Aggregation and Cross-Modality Interactive Learning for RGB-D-Based Human Action Recognition", *IEEE Access*, Vol. 10, pp.104190-104201, 2022.

76. A. Shahroudy, J. Liu, T.T. Ng, G. Wang, "NTU RGB+D: a large scale dataset for 3D human activity analysis", in *Proc. CVPR*, Las Vegas, NV, USA, pp. 1010–1019, 2016.

77. Gopampallikar Vinoda Reddy, Kongara Deepika, Lakshmanan Malliga, Duraivelu Hemanand,Chinnadurai Senthilkumar, Subburayalu Gopalakrishnan, and Yousef Farhaoui "Human Action Recognition Using Difference of Gaussian and Difference of Wavelet" BIG DATA MINING AND ANALYTICS ISSN 2096-0654 07/10 pp336 –346Volume6,Number 3, September 2023DOI: 10.26599/BDMA.2022.9020040 Presshttps://www.sciopen.com/article 2024

78. Vinoda Reddy, P.Suresh Varma,, A.Govardhan, "RECURRENT FEATURE GROUPING AND CLASSIFICATION MODEL FOR ACTION MODELPREDICTION IN CBMR" International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.7, No.5/6, November 2017.

79. Vinoda Reddy,, Dr.P.Suresh Varma,and Dr.A.Govardhan," MULTILINEAR KERNEL MAPPING FOR FEATURE DIMENSiON REDUCTION IN CONTENT BASED MULTIMEDIA RETRIEVAL SYSTEM", The International Journal of Multimedia & Its Applications (IJMA) Vol.8, No.2, April 2016

80. S. Swapna Rani N. Kalaivani, S V Hemanthm, P.N. Sundararajan, G. Vinoda Reddy, G.Amirthayogam,"Deep Learning based vehicle image detection using Yolo V5 with Region-Based Convolutional Neural Network", 2024 3rd IEEE International Conference on Artificial Intelligence for Internet of Things DOI: 10.1109/ICCAMS60113.20 23.10525942 (AIIoT 2024)

81. Vinoda Reddy, P Suresh Varma and A.Govardhan, "Recurrent Energy Coding for Content Based Multimedia Retrieval System", International Journal of Multimedia:User Design and User Experience, Vol. 23, No. 1, November 2015, pp. 1114-1120.

82. Vinoda Reddy, P. Suresh Varma and A. Govardhan, "Action model prediction and analysis for CBMR Application", Second IEEE Conference International Conference on Computing Methodologies and Communication (ICCMC2018), IEEEexplore ISBN-978-3452-3,3 March 2018, PP. 1015-1020 doi: 10.1109/ICCMC.2018.8487504ISBN- 978-3452-3,

83. Vinoda Reddy, P. Suresh Varma and A. Govardhan, "Intercorrelative Histogram Feature and Dimension Reduction for CBMR", International Conference on Data Mining (DMIN'16), Las Vegas, USA, pp. 24-30 .ISBN #: 1-60132-403-02-431-6 Proceedings of the International Conference on Data Mining DMIN16

1. X.Zhang,X.ZhangandL.Han,"AnenergyefficientInternetofThingsnetworkusingrestart artificial bee colony and wireless power transfer", *IEEE Access*, vol. 7, pp. 12686-12695, 2019. https://doi.org/10.1109/ACCESS.2019.2892798
2. X.Zhong,L.ZhangandY.Wei,"Dynamicload-balancingverticalcontrolforalarge-scalesoftware-defined Internet of Things", *IEEE Access*, vol. 7, pp. 140769-140780, 2019. https://doi.org/10.1109/ACCESS.2019.2943173
3. M. Malik, M. Dutta and J. Granjal, "A survey of key bootstrapping protocols based on public key cryptography in the Internet of Things", *IEEE Access*, vol. 7, pp. 27443-27464, 2019. https://doi.org/10.1109/ACCESS.2019.2900957