 doi:10.20944/preprints202005.0146.v1

Study on Machine Learning and Deep Learning Methods for Human Action Recognition

Gopika Rajendran[*1], Ojus Thomas Lee[1], Arya Gopi[1], Jais Jose[2], Neha Gautham[3]

[1] Computer Science and Engineering, A. P. J. Abdul Kalam Technological University, College of Engineering Kidangoor Kottayam, 686583 Kerala, India
[2]Amity Institute of Geoinformatics and Remote Sensing, Amity University, Noida, Uttar Pradesh
[3]University of Malaysia, Sarawak

*Corresponding Author: gopikaraj4@gmail.com

Abstract With the evolution of computing technology in many application like human robot interaction, human computer interaction and health-care system, 3D human body models and their dynamic motions has gained popularity. Human performance accompanies human body shapes and their relative motions. Research on human activity recognition is structured around how the complex movement of a human body is identified and analyzed. Vision based action recognition from video is such kind of tasks where actions are inferred by observing the complete set of action sequence performed by human. Many techniques have been revised over the recent decades in order to develop a robust as well as effective framework for action recognition. In this survey, we summarize recent advances in human action recognition, namely the machine learning approach, deep learning approach and evaluation of these approaches.

Keywords Human Action Recognition, Machine learning, Action feature Representation, Action Classification, Deep learning

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2020 doi:10.20944/preprints202005.0146.v1

1 Introduction

Analyzing human motion from a video is the most challenging task in wide-ranging applications such as computer vision as well as in computer graphics. One such application, particularly in computer animation, is the retargeting of motion from a performer to character to be animated. Since humans move in three dimensions and its video recordings are in two dimensions, we need to apply some 2D-to-3D pose recovery or pose estimation schemes. Retargeting approaches are applied only after the pose recovery schemes are evaluated. Research in pose recovery schemes leads to the area of the understanding human behavior so-called human action recognition (HAR). The ultimate goal of HAR is to build an intelligent machine that can accurately interpret the human behavior and actions from the video. The core of an intelligent machine is the computational algorithm that interprets human action. Equivalent to human vision system, the algorithm produces a label after observing the entire human action execution. Creating such algorithms is commonly viewed in computer vision research that considers how machine benefits from digital images and videos to high-level comprehension.

In context of computer vision, the term human action ranges from simple joint/bone movement to complex movement consist of multiple joint/bone(s) and human body. The human activity is dynamic in nature and usually communicated in a video enduring a couple of seconds. A human action endeavors to obtain certain goal, in which a few of them can be implemented essentially by basic action, while others have to be fulfilled in a few steps. Action recognition is an elementary task that recognizes human actions based on total action execution in a video by virtue of present state. An action in a video is depicted as an array representing pixel values in each frame. The machine has no idea regarding how to formulate feature information from these pixels that describes the action and how to infer human actions from these representations. So we can split the action recognition into problems of action representation and classification of action.

In this work, we outline most recent works and present a survey of researches in vision-based human action recognition techniques including action classification (especially on the action representation methods of action classification) and deep-learning methods.

The structure of paper is organized as follows: Section 2 defines a set of action recognition applications. Section 3 describes the main challenges in the domain of research. The basic approaches are described in Section 4 and Section 5.

Section 6 describes the databases used in these methods. Evaluation of these approaches is presented in Section 7. We conclude the paper by the Section **Error! Reference source not found.**.

2 Application of Human Action Recognition

There are many potential applications of human action recognition. In this section, we will review couple of applications such as Visual Surveillance, Human-Robot Interaction, Entertainment, Autonomous Driving Vehicle and Video Retrieval.

2.1 Visual Surveillance

In recent years, video surveillance [1][2] has earned great attention within the computer vision community. Astonishing desire for protection and security services has led to more challenging intelligent surveillance work. It has a wide variety of applications, from tracking human activities in public spaces such as waiting rooms, railway stations, clinics, nursing homes, campuses. Evaluation of behavior consists of analysis and identification of movement patterns to provide a high-level interpretation of actions and interactions among objects. Surveillance camera footage is most often used to track suspicious incidents or activities and is sometimes used to avoid potentially dangerous circumstances.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2020 doi:10.20944/preprints202005.0146.v1

2.2 Human-Robot Interaction

Robots [3] are utilized to perform a vast range of tasks, particularly in human environments. These environments make analysis of characteristic human-robotic interactions more significant. In several roles, such as personal indoor assistants, robots must be able to deduce human activities and determine whether or not human assistance is needed. Such an interaction calls for contact between robots and humans. Visual communication [4] is one of the cleverest ways of applying such communication.

2.3 Entertainment

The gaming industry has accumulated a vast and diverse bunch of people by introducing different kinds of gaming consoles. A new wave of games focused on full body motion centered play like dance and sports games increased the gaming platform's attractiveness. To allow accurate perception of human behavior, these gaming consoles provide low-cost RGB-Depth sensors such as kinect camera [5] that provide information on the human skeleton in detail. This promotes the action recognition function, which optimizes intra-class activity variations and reduces cluttered background noise.

2.4 Autonomous Driving Vehicle

Recognition of human behavior within vehicles are becoming increasingly important due to the successful mapping of driver and passenger activities [7] in autonomous vehicle. With the accurate classification about what the driver is doing, the human machine interaction can be directed to the suitable modalities. If the car knows the full body movement of its passengers, the safety functions can be adapted to the in-the-moment deployment for example braking, airbag, crash avoidance etc. In-vehicle activities and driver attention in the driving task can be used by the autonomous vehicle (AV) to decide whether perform a safe stop. A major concern about fully automated vehicles is that it should provide a design of the interaction between the user and the automatic system or vehicle.

2.5 Video Retrieval

At present, due to the rapid growth of technology, the number of videos available on the Internet is increasing significantly. Videos have become popular with widely available equipment that can make, upload, and share images, high-speed Internet access, and free storage servers. Due to the huge amount of video data being uploaded to the Web every second, the viewing of videos and movies on the Web seems to be quite prevalent today. Such a great deal of video data and its popularity have led to video understanding and action recognition for the retrieval of relevant videos. Content-based video retrieval [8] is used for finding the user desired items among these big video data. Majority of the videos are related to the human actions. So retrieval of human action based videos from the big data of videos is becoming more popular.

3 Challenges in Research

In this section, we will define the domain of challenges that are experienced while considering the section 2 such as Intra-class Variation and Inter-class Similarity, Complex and Various Backgrounds, Uneven Predictability.

3.1 Intra-class Variation and Inter-class Similarity

As we can see, human may behave differently for the same activity. For example, running activity may vary by the speed of execution even with intermittent jump. In most expressions, one category of behavior can include many different types of human movements. Furthermore, different people can present different poses in the same action. Such actions results in intra-class variation which can confuse the existing action recognition algorithm. At the other hand, different practices can express similar forms known as inter-class similarity. Such scenarios can be observed from running and walking activities which involve similar motion pattern which leads to miss-classification. Such problems must be solved by accurate and distinctive features derived from videos.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2020 doi:10.20944/preprints202005.0146.v1

Complex and Varying Backgrounds

Varieties of algorithms for human action recognition function well in indoor controlled environments but not in outdoor environments. Such deviations occur because these videos are complex in nature and vary with the environment and therefore introduce background noise. Most of the current action features such as histogram of oriented gradients encounter background noise that downgrades the output of the recognition. Another factor that can be used in real world applications is camera movement. Due to the significant camera movement the action function cannot be effectively extracted. To solve this, camera motion should be sculpted and taken care of.

3.2 Uneven Predictability

An action can be represented in the video with a number of frames. Some of these frames do not represent significant information and contain some outliers. Hence action prediction methods require that the beginning portions of the video need to be informative in order to maximize predictability. Such problems can be solved by transferring the context information to the beginning portion of the video streams.

4 Machine Learning Approach

Traditional recognition of actions usually comprises two key components: description of actions (action representation) and classification of actions. The action representation translates an action video into a function vector or a set of vectors, and action classification implies an action mark/label from that vector. Action recognition algorithm takes an image or a patch of an image as input and outputs the context of the images. The interpretation of human actions in video action not only determines the presence of the person(s) in the image space, but also includes descriptions of emotional interaction, identity, gender, etc. Figure 1 shows the anatomy of an action classifier. Before machine learns the classification, the input to the machine must be made acceptable with minimum classification error through a process called Preprocessing. The generated image sequence is usually well-segmented and comprises only one action event and therefore becomes a classification problem.

4.1 Action Representation

The foremost problem about recognition is to depict an action in a video. Human acts that appear in videos vary due to their speed of motion, camera view, posture and pose variations, making it a very challenging issue. A good method of representing actions should be accurate for estimating and characterizing actions effectively. The goal of the action representation is to turn an action video into a feature vector that extracts representative and biased knowledge about human actions and minimizes variations and thus increases the recognition efficiency. The main problem of action representation is expanded from 2D (related with space domain) to 3D (related with space-time domain). The approaches to action representation can be classified into global features and local features based on spatial and temporal variations. Such features are used mainly to define behavior in traditional machine learning methods, such as support vector machines and probability map models.

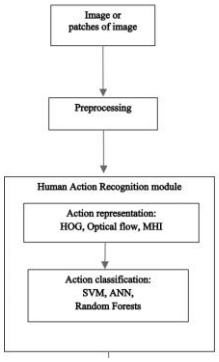

Figure 1: Anatomy of machine learning based action classifier

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2020 doi:10.20944/preprints202005.0146.v1

1. ***Local Representation:*** Local Representation identifies the local regions having salient motion information. The key benefit of the local features is that there is no need for knowledge about human body model or people localization. Local features [8] are extracted by applying a local feature detector or descriptors of local features, and then encoding spatiotemporal neighborhoods around the features identified using a local feature descriptor. Local feature capture shape as well as details on motion in a local neighborhood surrounding points of interest and trajectories. One of such most popular descriptors are HoG features. In [10], Seemanthini et al., the gradient orientation are computed in which they occur in a coined region of an image. At many image regions, patches of multiple scaled image are analyzed by the gradient histogram algorithm. They are designed to offer robustness for local appearance and position changes. Edge gradients and orientations are calculated at each pixel within the given local area in order to measure the HoG features.

Takuya et al. [11] propose a Sobel filter used to get gradients and orientations of edges. The magnitude and orientation of the gradients was determined using the spatial x and y gradients. The local area that surrounds a local feature is then separated into a spatio-temporal grid that includes cells. A histogram descriptor is determined for each cell of the grid. Thus forming final descriptors by normalizing and concatenating the histograms per cell in a given frame or image. The final descriptors constitute of a HoG features vector. The HoG function vector defines the local shape of an object in plural cells, with edge information. While in flat regions such as a floor or a house wall, the histogram of the oriented gradients has featureless distribution. Other important descriptors used in an action recognition algorithm are an optical flow histogram. The HOF descriptor encrypts information related to motion and speed. The optical flow, Silambarasi et al. [12], is calculated over segmented human objects. The human object from the boundary detection method is considered for optical flow extraction. The optical flow vector has horizontal and vertical component at each pixel and quantized into the histogram bins. HoF is more applicable when actions are recognized with similar shape but varying speed. Silambarasi et al. proposed a fusion of HoG and bag of HoF because HoG performs better when the gradient or shape is sufficient to recognize.

2. ***Global Representation:*** Global Representation directly extracts and represents global descriptors from original videos or images as a whole. The human subject is localized and extracted in this representation using methods of background subtraction by framing the silhouettes or shapes, i.e., Region of Interest (ROI). ROI are encoded as descriptors in which edges, corners, or optical flows are derived.

An adaptive idea is to extract the human body from the background so that human actions can be identified in videos. Such an extraction technique is called background or foreground extraction. The extracted shape is called silhouette, which in the global representation approach would be the region of interest and represented as a whole entity. Combining silhouette over all frames generates motion history image about the activities described in the video. The motion history image (MHI) is a static image representation that illustrates movement direction and path as it evolves. In MHI, the information about temporal motion is broken down into a single image structure as in Figure 2, where function of motion is termed as intensity. In [14], MHI can be obtained by extracting a binary image sequence that indicates the motion regions by means of frame differencing. Based on binary image sequence, the motion history image is defined as shown in Figure 2. The resultant image is a gray-scale image where brighter pixels represent the motion. The MHI is called a function vector for defining action in the image sequence. Extract the feature vectors of all action sequences during training, which are identified as temporal template, and extract the feature vector of an unknown action sequence in the recognition process. In the next step, the function vector is compared with the temporal template resulting in a classified mark/label. Another method for describing high-level motions of interest, MHI-based regions, is detected as 3D volume space in the frame.

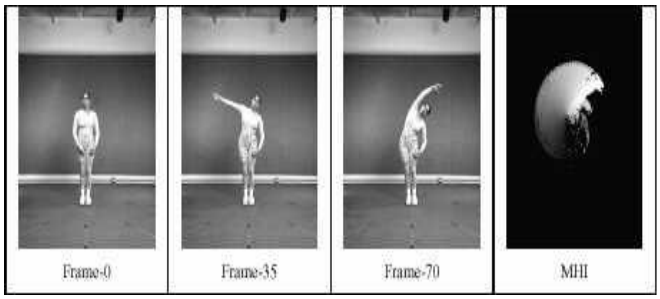

Figure 2: Motion history Image created from image differences for arm-stretching movement [13]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2020 doi:10.20944/preprints202005.0146.v1

4.2 Action Classification

After computing the feature vector, action classifiers should be learned from training samples such that the class boundaries for specific action classes are calculated. In this point, the activity label is given as a final output by the classification algorithm. Some of the classification algorithms are:

1. ***Support Vector Machine (SVM):*** SVM is an informatics framework over a series of similar supervised learning methods that analyze data and identify patterns of classification and regression. The standard SVM takes a set of input data and decides if it belongs to any of the categories for a given input. Therefore SVM becomes a non-probabilistic binary classifier. Based on a set of training samples, SVM training algorithms, used in [12],[14],[15], build a model which assigns new examples to one category or another. Multiclass SVMs [16] are made using SVM to assign labels to a case, so that the labels are drawn from a finite set of multiple samples. An effective approach to do so is to reduce the multi-class problem into several problems of binary classification. In this way, Multi-class SVM trains the framework which will give the category of action with respect to the features extracted.

2. ***Random Forests:*** Random forests or Random decision forests are the method of learning classification, regression and other tasks that function in creating a multiple decision tree at the time of training and class performance. It can be a model of classes (classification) or mean prediction (regression) of the individual trees. The generated feature vector is used to control the tree's growth. In [17], Lu Xu et al utilities ID3 algorithm for constructing random forest consisting of k tree classifiers using accelerometer data generated from sensor, as in Algorithm 1. During the growth of the tree, randomly select n dimension at each node which defines the tree's growth. Then select the best dimension for splitting the training set on the tree to expand. As each node is divided into child nodes, the training set on the current node is divided into respective child nodes. After creating random forest, they vote at the feedback for the best class. For this purpose, they defined the algorithm as in.

```
function Random_forest(data,m) return Rand
Input     : data, training set
            m, the number of tree
Output    : Classifier
Initialization:
    Rand ← m decision tree
    train_set ← obtain m sample set on data
    target_set ← set of train data
    for i = 0 to m do
        decision tree ← ID3(train_set[i],
        target_set[i]);
        Rand.add(decision tree)
```

Algorithm 1: Random Forest

Chi-Hung Chuan etc al [5], proposes another concept of random forest called action forest. It uses decision function instead of test conditions at each node were decision function holds practical physical means. The same algorithm with decision function can be used for this purpose. The splitting of training set is defined over decision function. Then it calculates the distribution for all leaf nodes of each class and selects the highest likelihood as the vote. After construction of the decision tree, each leaf node contains some dataset and classified result. The voting item is extracted to boost the accuracy and efficiency of each leaf node by maximizing the amount of the classified result with the number of data. The category outcome is a cumulative vote of all the decision trees and the largest number of votes in the category.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2020 doi:10.20944/preprints202005.0146.v1

5 Deep Learning Approach

Recent technique uses deep networks that combine the component of action representation and the component of action classification into a single end-to-end trainable frame work, and further enhance overall classification performance. Due to their ability to build a powerful feature which can be easily generalized, features learning using deep learning have gained great attention in recent years. The performance of deep learning in the identification of action has contributed to network growth from ten to millions of parameters and a large labeled dataset. While developing a deep network, the two major variables that should be defined are convolution operation and temporal modeling. Convolution operation is the fundamental component of deep networks which aggregate the pixel value in a small spatio-temporal neighborhood using mask (or kernel matrix). It is also known as the filter with leavable parameters used to extract low dimensional characteristics from the input data. There are 2D convolution and 3D convolution. 2D convolution is the convolutional filter that moves in two direction (x,y) to calculate the low dimensional features from the data. Output will be the 2D matrix. Whereas 3D convolution is the convolutional filter that moves in three dimension (x,y,z) and compute the low feature representation.

Temporal modeling in deep networks captures the temporal information from the consecutive frames. There are three temporal modeling approaches. One approach is to apply 3D convolution directly to several consecutive frames which will capture the temporal dynamics in those frames as the temporal component in the 3D convolution process. Other approach is to model using multiple streams. Streams that indicate flow in network are trained on optical flow frames which capture the temporal details. The third category of approaches uses temporary aggregation or pooling to capture temporal information in a video. Convolutional Neural Network (CNN) is the simple deep networks which are used in action recognition. CNN is programmed primarily to derive spatial 2D characteristics from still images. Since videos are generally interpreted as 3D spatio-temporal signals, Guangle Yao et al,[20]extends the CNN from image to video, using temporal information. It is composed of an input layer, an output layer and several hidden layers. The layers concealed are convolutional, pooled or fully connected. Convolutionary layer performs an action of convolution and an additive bias on the raw data and passes the effect first via an activation function and then into the next layer. Most CNN extracts the feature map or heat map which explains the probability of joint occurrences in both 2D and 3D. To estimate the 3D coordinates, the network must infer the 2D coordinates first. Using the estimated 3D coordination of joints over the entire video sequence, actions are recognized.Wen-Nun et al. [21] introduced two fully convolutional neural networks that infer the 3D pose from 2D features of a given image or video. In [21,[23], construct recurrent neural network (RNN) or CNN-based networks with long-term memory units (LSTM) [24]. The trained model should pick discriminative skeletal joints within each input frame and pay greater attention to specific frame outputs. As the architecture grows higher, the training time has a negative impact on the performance. As a solution to this problem, Francesc [25] has formulated the 3D action recognition problem as a 2D-3D distance matrix regression in which the data dimension is reduced to a small one. Therefore, mapping from input to output can be accomplished across shallow architectures. Using Fully Connected (FConn) and Fully Convolutional (FConv) networks, we can learn a regression function which maps 2D to 3D distance matrix.

Table 1: Comparison of Machine Learning Approach

Sl no.	Reference	Input Data	Assumption	Features	Classifier
1	Lu Xn et al.[17]	Skeleton Data	-	-	Random Forest
2	Jamie Shotton et al.[18]	Depth Data	-	Coded body features	Random Forest
3	Chi-Hung Chuan et al [5]	Skeleton Data	-	-	Action Forest
4	ya Kostrikov [19]	RGB-image	Hypothesis a plane for depth information	set of 2D images with 3D annotated data	Regression Forests
5	Chin-Pan[14]	Video	-	HoG of MHI	SVM
6	Silambarasi[15]	Video	-	HoG of MHI	Multi-class SVM
7	Silambarasi [11]	Video	-	HoG + Bag of HOF	SVM

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2020 doi:10.20944/preprints202005.0146.v1

6 Datasets

The database used in action recognition can be categorized as controlled action video dataset and uncontrolled action video dataset depending on the environment recorded by the camera.

6.1 KTH dataset

KTH database [27] includes 6 forms of human behavior (Walking, jogging, jumping, punching, hand waving and hand clapping) carried out many times by 25 subjects in four separate scenarios: outdoor (s1), outdoor with scaled variations (s2), outdoor with different clothes (s3) and indoor (s4) combinations. This falls in the category of controlled video action dataset. All such frames were taken over from homogeneous backgrounds with a static camera with a frame rate of 25fps.

6.2 HumanEva-I Datasets

The HumanEva-I [28] dataset is the guided action dataset, which includes 7 highly optimized sequences (4 grayscale and 3 color) coupled with 3D body motions from a motion capture system. It contains 4 subjects that perform 6 traditional activities such as walking, jogging, gesturing, etc. Participants are given the 2D and 3D pose error metrics for measuring errors. The dataset includes training sets, evaluation and testing of ground-truth or references.

6.3 Human3.6M dataset

The Human3.6M dataset [29][29] is the compilation of 3.6 million accurate human poses collected from four different viewpoints by capturing the output of 5 female and 6 male subjects. It is used to train realistic human sensing systems and test the next step of human pose model and algorithm predictions. Then it offers direct mixed reality evaluation scenarios in which 3D human models are animated with motion capture and integrated using appropriate 3D dynamics.

6.4 NTU RGB+D Dataset

NTU RGB+D dataset [30] is a large-scale RGB+D human action recognition dataset with over 56,000 video samples and 4 million frames from 40 different subjects. The dataset includes 60 different types of behavior including regular, mutual, and health related acts.

7 Evaluation of Machine Learning and Deep Learning approach

In this section we present an evaluation of the methods we surveyed based on the findings in the literature. MPJPE, accuracy and JPE were used as parameters for the evaluation of the methods.Machine Learning approaches are evaluated on small-scale datasets as KTH dataset as in section 4. Cross validation training schemes are used in machine learning methods and confusion matrix is used to demonstrate the accuracy of recognition over each category of action. A confusion matrix provides a description about the predictive results in a classification problem. Accuracy is evaluated in the terms of confusion matrix feature. Table 1 provides the comparison of some literature based on machine learning and its performance result. The accuracy values for the machine learning method are shown in Figure 3. From the plot it is clear that the approach proposed by CHi Hung et al. gives maximum accuracy. The high accuracy rate is due to multiple features extracted in a image or video performs better than extracting single features since they suffer from background noise.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2020 doi:10.20944/preprints202005.0146.v1

Table 2: Comparison of Deep Learning Methods

Sl no.	Reference	Input Data	Architecture	Dataset	
1	Wen-Nung[20]	RGB-Image	3 CNN	Human3.6M	
2	Francesc Moreno[24]	RGB-Image	FConn + FConv	Humaneva-I, Human3.6M	
3	Yu Liu et al. [25]	RGB-Image	CNN + EM	MOCAP	
4	Charissa Ann[21]	Sensor data	CNN + LSTM, RNN + LSTM	UCI dataset	
5	Earnest[26]	Video		CNN	UCF50 dataset
6	Yinghao Huang[23]	Sensor data		LSTM	AMASS dataset
7	Hany El-Ghaish [22]	3D Skeleton data, Body part image, MHI		2 CNN + LSTM, CNN	NTU RGB-D Dataset

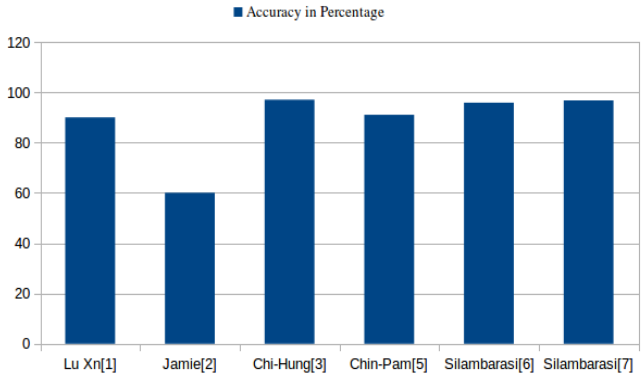

Figure 3: Accuracy values observed in the methods as in Table 1 using confusion matrix.

Deep learning networks are typically tested on large-scale datasets such as Human3.6 M and show overall performance of recognition on each dataset. There are several metrics for evaluating. Most popular metrics used for action recognition is Mean per joint position error (MPJPE) and joint position error (JPE). JPE is the Euclidean distance between ground-truth or reference feature and prediction for a joint. MPJPE is the mean of JPE for all N joints features. Lower MPJPE value shows better performance. MPJPE is defined in the equation **Error! Reference source not found.** .

$$MPJPE = {}^{1}/_{T}\,{}^{1}/_{N} \sum_{t=1}^{T} \sum_{n=1}^{N} \left\| \left(j_n^{(t)} - j_{root}^{(t)}\right) - \left(j_n^{(\vec{t})} - j_{root}^{(\vec{t})}\right) \right\|_2 \quad (1)$$

MPJPE is measured only after the estimated root joints and the ground-truth of 3D pose are matched. The joints are also standardized for the root joint.[31,32] The goal of this metric is to achieve high accuracy in recognition in order to identify action as accurately as possible. Error! Reference source not found. explains the various designs and performance of architectures in deep networks. From the Error! Reference source not found., it is evident that simple architecture performs better than complex architectures and machine learning approaches. Datasets has an impact on the performance where MPJPE values are defined. As a result, deep networks with simple architecture and less MPJPE values are best suited for tasks like human pose estimation and motion retargeting to detect behavior.[33,34]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 9 May 2020 doi:10.20944/preprints202005.0146.v1

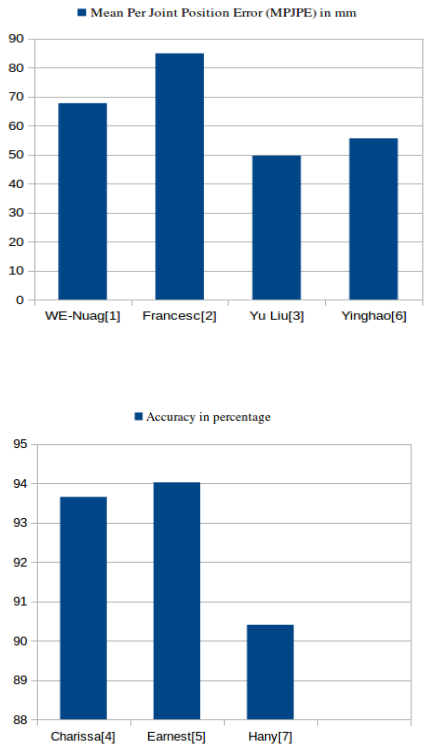

Figure 4: a) MPJPE values and b) Accuracy values describes the performance of the deep learning approaches as in Error! Reference source not found..

8.Conclusion

The explosion of big data related videos and the recent developments in computer vision divert the research focus from understanding human action to human action perception. We presented a survey of the cutting-edge technologies in human action recognition. Within this paper we discuss the latest trends within motion recognition including application, issues of action recognition, basic approaches and their assessment. Deep networks have the superior performance in learning problems with the action feature. Deep learning networks are capable of extracting complex data structures rather than simply action labels, make them a dominant approach in action recognition schemes.

Reference

[1] Ahmed Taha et al. “Human activity recognition for surveillance applications”. In: Proceedings of the 7th International Conference on Information Technology. 2015, pp. 577–586W. Zabierowski, A. Napieralski. Chords classification in tonal music, Journal of Environment Studies, Vol.10, No.5, 50-53.

[2] Michal Koperski. “Human action recognition in videos with local representation”. PhD thesis. 2017.

[3] Addwiteey Chrungoo, SS Manimaran, and Balaraman Ravindran. “Activity recognition for natural human robot interaction”. In: International Conference on Social Robotics. Springer. 2014, pp. 84–94.

[4] David Vogt et al. “Learning human-robot interactions from human-human demonstrations (with applications in lego rocket assembly)”. In: 2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids). IEEE. 2016, pp. 142–143.

[5] Chi-Hung Chuan, Ying-Nong Chen, and Kuo-Chin Fan. “Human action recognition based on action forests model using kinect camera”. In: 2016 30th International Conference on Advanced Information Networking and Applications Workshops (WAINA). IEEE. 2016, pp. 914–917. Cristofer Englund and Martin Torstensson. “In-vehicle Driver and Passenger Activity Recognition”. In: SSBA. 2019.

[6] Cristofer Englund and Martin Torstensson. “In-vehicle Driver and Passenger Activity Recognition”. In: SSBA. 2019.

[7] Mohsen Ramezani and Farzin Yaghmaee. “A review on human action analysis in videos for retrieval applications”. In: Artificial Intelligence Review 46.4 (2016), pp. 485–514.

[8] Piotr Tadeusz Bili´nski. “Human action recognition in videos”. PhD thesis. Universit´e Nice Sophia Antipolis, 2014.

[9] K Seemanthini and SS Manjunath. “Human Detection and Tracking using HOG for Action Recognition”. In: Procedia computer science 132 (2018), pp. 1317–1326.

[10] Takuya Kobayashi, Akinori Hidaka, and Takio Kurita. “Selection of histograms of oriented gradients features for pedestrian detection”. In: International conference on neural information processing. Springer. 2007, pp. 598–607.

[11] Suraj Prakash Sahoo, R Silambarasi, and Samit Ari. “Fusion of histogram based features for Human Action Recognition”. In: 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS). IEEE. 2019, pp. 1012–1016

[12] Francisco Mora Lizfffdfffdn et al. “Working with OpenCV and Intel Image Proccessing Libraries. Proccessing image data tools”. In: (July 2002).

[13] Chin-Pan Huang et al. “Human action recognition using histogram of oriented gradient of motion history image”. In: 2011 First International Conference on Instrumentation, Measurement, Computer, Communication and Control. IEEE. 2011, pp. 353–356.

[14] R Silambarasi, Suraj Prakash Sahoo, and Samit Ari. “3d spatial-temporal view based motion tracing in human action recognition”. In: 2017 International Conference on Communication and Signal Processing (ICCSP). IEEE. 2017, pp. 1833–1837.

[15] VK Swarnambigai. “Action Recognition Using AMI and Support Vector Machine”. In: IJCST 5.4 (2014).

[16] Lu Xu et al. “Human activity recognition based on random forests”. In: 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD). IEEE. 2017, pp. 548–553.

[17] Jamie Shotton et al. “Real-time human pose recognition in parts from single depth images”. In: CVPR 2011. Ieee. 2011, pp. 1297–1304.

[18] Ilya Kostrikov and Juergen Gall. “Depth Sweep Regression Forests for Estimating 3D Human Pose from Images.” In: BMVC. Vol. 1. 2. 2014, p. 5.

[19] Guangle Yao, Tao Lei, and Jiandan Zhong. “A review of Convolutional-Neural-Network-based action recognition”. In: Pattern Recognition Letters 118 (2019), pp. 14–22.

[20] Wen-Nung Lie et al. “Fully Convolutional Network for 3D Human Skeleton Estimation from a Single View for Action Analysis”. In: 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE. 2019, pp. 1–6.

[21] Charissa Ann Ronao and Sung-Bae Cho. “Human activity recognition with smartphone sensors using deep learning neural networks”. In: Expert systems with applications 59 (2016), pp. 235–244.

[22] Hany El-Ghaish et al. “Human Action Recognition Based on Integrating Body Pose, Part Shape, and Motion”. In: IEEE Access 6 (2018), pp. 49040–49055.

[23] Yinghao Huang et al. “Deep inertial poser: learning to reconstruct human pose from sparse inertial measurements in real time”. In: SIGGRAPH Asia 2018 Technical Papers. ACM. 2018, p. 185.

[24] Francesc Moreno-Noguer. “3d human pose estimation from a single image via distance matrix regression”. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017, pp. 2823–2832.

[25] Yu Liu et al. “End-to-End Algorithm for Recovering Human 3D Model from Monocular Images”. In: 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). IEEE. 2018, pp. 1082–1087.

[26] Earnest Paul Ijjina and C Krishna Mohan. “Human action recognition based on recognition of linear patterns in action bank features using convolutional neural networks”. In: 2014 13th International Conference on Machine Learning and Applications. IEEE. 2014, pp. 178–182.

[27] Barbara Caputo Ivan Laptev. Recognition of human actions-KTH dataset. Jan. 2005. URL: http://www.nada.kth.se/cvap/actions/.

[28] Leonid Sigal, Alexandru O Balan, and Michael J Black. “Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion”. In: International journal of computer vision 87.1-2 (2010), p. 4.

[29] Catalin Ionescu et al. “Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments”. In: IEEE transactions on pattern analysis and machine intelligence 36.7 (2013), pp. 1325–1339.

[30] Amir Shahroudy et al. “NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis”. In: IEEE Conference on Computer Vision and Pattern Recognition. June 2016.

[31] Nathalia, Deepa, Arjun Suresh, and Neha Singh. “Monitoring Land Use / Cover Changes during the Mining Activities in Aravalli Hill Region.” 3.6 (2018): n. pag. Print.

[32] jose, J., Yuvaraj E., Kushik B., Neha Singh., Suresh A., Global Tsunami Hazard Web Map For Mitigation And Rescue Operation. International journal of scientific & technology research. (2020) volume 9, issue 01

[33] Bhardwaj, Pallavi et al. “Satellite Monitoring for Spatio-Temporal Changes Occurring in Forest Area of Sariska Tiger Reserve by Implementing GIS and Remote Sensing Techniques.” 10 (2019): 26–36. Print.

[34] Jose, J.; Yuvaraj, E.; Aswin, S.; Suresh, A. Development of Worldwide Tsunami Hazard Map for Evacuation Planning and Rescue Operations. Preprints 2020, 2020040370 (doi: 10.20944/preprints202004.0370.v1).

 doi:10.20944/preprints202005.0146.v1