Chapter 2 THEORETICAL BACKGROUND AND LITERATURE REVIEW

2.1. Introduction

Human activities can be seen as a sequence of basic motions. For example, activities like brushing hair or hand waving can be described as a sequence of consecutive raising and lowering of the hand. There are different sorts of human activities [4] according to their complexity; these activities are divided into four types: “Gestures”, “Actions”, “Interactions”, and “Group Activities”. “Gestures” are the simple motion of a part of a body. For example, “raising an arm and moving a leg”. Actions will be exercises performed by one individual that might be made out of various gestures organized in a time order, “strolling”, “waving”, and “punching” are examples of “Actions”. “Interactions” are human activities that involve at least two individuals or objects. As an example, “two persons checking hands” is an interaction between two individuals and “somebody pushing table” is an interaction that includes one person and an object. Finally, “Group activities” are that activities played by a group composed of individuals or objects: “A group of people playing football” and two groups fighting” are typical examples.

Monitoring of changes in an actor's behaviour is an important process in activity recognition. This task is in charge of acquiring applicable relevant data for activity recognition systems to recognize an activity. There are several input modalities which can be used for activity recognition tasks, e.g. vision-based video frames, depth map (Fig. 2.1), skeleton etc. A Vision-based activity recognition utilizes computer vision methodologies to analyse visual observations for activity recognition using visual sensing facilities, e.g., camera, and infra-red sensor, to capture activity. There has been significant work made on vision-based activity recognition [5], however, because of the multifaceted nature of true settings. These methodologies experience issues related to reusability and scalability, such as highly variation of activities in the natural environment. “Depth Maps-based” activity recognition depends mostly on features, either local or global, extricated from depth map images [5]. Depth maps give metric estimations of the geometry while visual information gives projective one that is invariant to lighting. Depth-map human activity recognition can be considered in its simplest form as a sequence of image representation, feature extraction process and recognition of these activities.

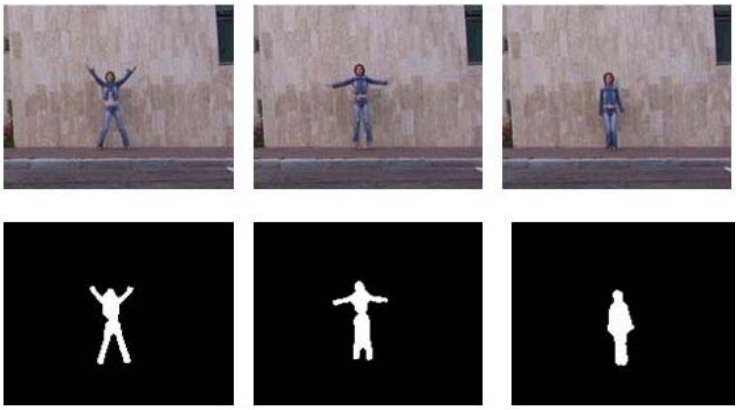

Figure 2.1 RGB frame in row 1 and depth frame in row 2

In the context of action recognition, a good feature representation must “be easy to compute”, provide a description for a sufficiently large class of actions”, “reflect the similarity between two like actions” and “be robust to various variations (e.g., view-point, illumination)”. Activity recognition solutions can be categorised into two types, namely representation based solution and deep learning based solution (Fig 2.2). Representation based approaches to human activity recognition follow the conventional approach, which consists of two steps: The first step computes complex handcrafted features from raw video frames and the second step learns classifier model or decision function based on the features obtained in the earlier step given labelled training video dataset. But it is hardly known which features are important for the recognition task at hand because the choice of features is highly problem-specific. For human action recognition, this is even more difficult since different activity classes may vary dramatically in terms of appearance and motion patterns involved. Deep network models on other hand are able to learn a hierarchy of features by building high-level, more complex features from low-level, simpler ones, thus automating the process of feature extraction. Hence, deep models do not require domain knowledge and a heavy burden of labour intensive feature engineering. Hierarchical nature of deep models also makes it capable of processing huge volumes of image or video data by computing more abstract concepts in terms of less abstract ones.

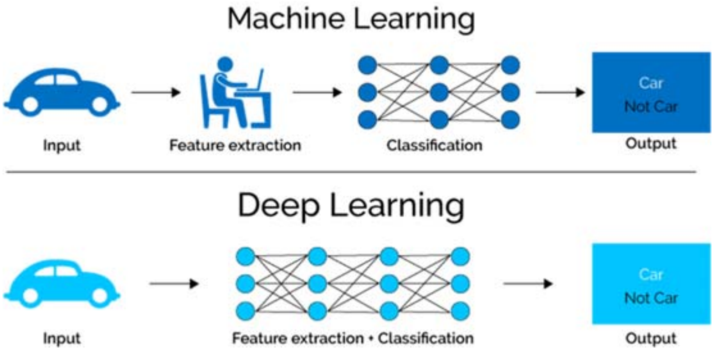

Figure 2.2 Activity Recognition Approaches [6]

In this chapter, a detailed literature survey has been presented, which includes a survey of various approaches for Human Activity Recognition and a survey on the evolution of modern datasets for human activity recognition. Literature review of human activity recognition approaches has been further accomplished in two parts - conventional/ hand-crafted features based approaches and deep learning based approaches. Moreover, the hierarchy of different approaches under them have been discussed, and research gaps have been identified. Further, we have done a comprehensive literature survey on the evolution of modern datasets for HAR. With the evolution of the human activity recognition field, the datasets used for training and testing of the proposed models have also undergone considerable change. In our survey, we attempted to classify and describe a verity of datasets for researchers to choose the most suitable benchmark for their domain. A set of characteristics have been proposed by which datasets may be compared. Finally, a detailed list of the dataset used for training and evaluation of proposed models have been presented, followed by a discussion on performance measures used in this thesis.

2.2. Literature Review of Human Activity Recognition Approaches

Over the past decade, a great deal of works has been done on the recognition of human activities. In literature, different kind of methods exists for activity recognition. This section presents a detailed survey of various human activity recognition frameworks under two categories, namely conventional machine learning based approaches and deep learning based approaches.

2.2.1. Conventional Machine Learning Based Approaches

The first category for Human Activity Recognition methods is a conventional approach also known as handcrafted feature based approach and is very much popular among human activity recognition community from past decades and also has shown very interesting results among well-recognized datasets. In this approach, comprehensive features are collected from the frames of images and based on that the feature classification is done using classifier such as Support vector machine. The brief classification of handcrafted features is shown in Figure-2.3.

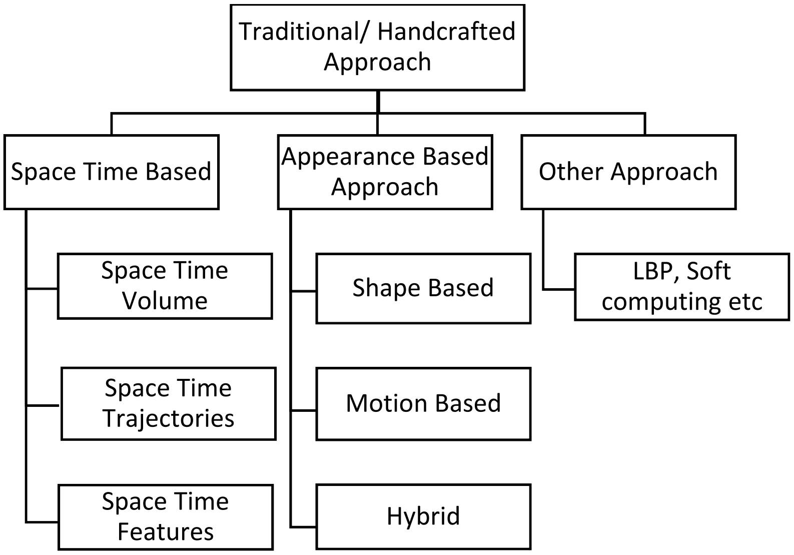

Figure 2.3 Conventional Human Activity Recognition Approaches.

2.2.1.1. Space Time Based Approach

In Space Time based approach, local and global features are extracted from the input video using sparse and dense feature extractor after that aggregation of methods and vocabulary is done which further is used for classification of action using supervised and unsupervised learning methods. Soumitra *et al.* [7] proposed a facet model using Space Time Interest Points (STIP). STIP is detected using 3D Facet Model than these points are represented using 3D Haar wavelet transform. Vocabulary is learned for 200000 descriptors for classification purpose using SVM as classifier. Kourosh *et al.*[8] presented the approach in which local space time features are extracted using Harris detector algorithm and histogram of optical flow is used as a descriptor. Also, Least Square Twin SVM is used instead of SVM for classification purpose, which makes the classification process four times faster. Shillin *et al.* [9] proposed the new method for calculation of STIP using Harris operator in the time dimension. First of all, a smoothening filter is applied for lightening and speed variations of actions in videos afterwards bag of words and bag of features are computed and are used by radial basis function in adaboost SVM for classification purpose. Instead of using STIP ShugaoMa *et* al. [10] proposed spatiotemporal space trees for classification directly from training data. Using the structure of the discovered tree and simple action words action classifier is made and SVM is used for classification purpose. Tree structure captures space, time and hierarchical relationships among the action words for classification. Tian *et al.*[11] put forward the concept of mid-level action elements. The input video is automatically segmented into mid-level elements according to spatiotemporal hierarchy than these are learned using a segmentation tree, and also action labels are provided with the weakly supervised setting. Chao *et al.*[12] presented an approach in which the max space-time graph is constructed for each activity class and afterwards subgraph with the max score is used for activity classification. The method is robust for detection in noisy backgrounds and also preserves relationships between objects and humans. Eleonora *et al.*[13] proposed an approach in which space-variant saliency-based processing by using different saliency models are used such as simple, central mask, single parameter and many more. Informative regions are extracted by using saliency-mapping algorithms and descriptors corresponding to them are exclusively done or by adding extra weight vector. Also, eye movements are used for exploration of the motion saliency approach. Chunyu *et al.*[14] put forward the approach in which location of human joints are estimated and are grouped into five body parts further data mining techniques are used to mine spatial-temporal pose structures for action representation. Also, spatial parts set corresponding to configurations of body parts and body part movements corresponding to temporal-part-sets are extracted in one frame, which is characteristic for human actions.

Heng *et al.*[15] proposed the method for extraction of dense feature trajectories by tracing the dense points of the optical flow containing the motion information from videos. The method is robust to camera stabilization, lighting and varying scenes also method outperform when compared to a bag of features approach. Xiaojiang *et al.*[16] suggested dense trajectory motion boundary-based sampling strategy which considers trajectories which are on the motion boundary. Moreover, motion boundary is generated using optical flow and regions without motion boundary foregrounds are deleted; also, new co-occurrence descriptor consisting of static and appearance information is used for activity recognition. Yu *et al.*[17] further extended the approach of dense trajectories and fused it with local and global motion points which leads to methodology in which no separation of background takes place while doing action recognition directly features are extracted, and dense trajectories are recognized. Mihir *et al.*[18] proposed the method in which visual motion is decomposed into space time trajectories further used by DCS descriptor which is based on shear information, divergence and curl. Also, VLAD coding methodology is used for recognition of actions purpose, which makes this method well capable of performing complimentary.

2.2.1.2. Appearance Based Approach

Appearance based approach is the approach in which physical structure is used in recognition purpose, and this approach is further classified into three subcategories, namely shape based, motion based and hybrid. Some of the prominent works available in literature corresponding to this category are discussed below. Shape based action recognition uses postures which are used to recognize human actions directly from the video. Basura *et al.*[19] proposed the Video Darwin method based on the temporary ordering of the frames from a video according to the appearance information and then using them for learning of ranking machine. Information is not captured only according to appearance but also over the evolution of time, and SVM is used as a classifier. Liang *et al.*[20] presented an approach in which fusion of feature extraction by kernel principal component analysis and factorial conditional random field (FCRF) based motion modeling are combined, making this model suitable for other temporal data analysis. Haiyong *et al.*[21] experimented method for directly recognition of actions from input video. Postures are recognized by Radon transform according to shape and edit distance as a classifier. Radon transform comes with lots of features such as low computational capacity, robust to frame loss and various drills in shapes. Nazli et al.[22] presented a newer approach Histogram of Oriented Rectangle (HOR) in which whole body is represented in the form of candidate rectangles. Then these spatial rectangles captured are used to form human silhouettes using different classification techniques such as SVM, Nearest neighbor classifiers, Motion Energy image and many more. Feng *et al.*[23] presented the approach in which shape and motion features are combined which uses multiple vision angles. The motion of human body is represented by 128-dimensional optical flow extracted from Region of interest and eigen shape vector of 90 dimensional is used to represent body shape also HMM are used to represent the activity set. Meng *et al.*[24] proposed a mid-level representation technique of Lie Algebraized Gaussians (LAG) in which spatio temporal features such as HOG3D are pinched at different scales and human bounding boxes are used for making video specific Gaussian Mixture Model which are further used to vectorize LAG for SVM classification of actions. Xinxiao *et al.*[25] presented an approach in which multiple XYT regions from training videos are used to learn Global Gaussian Mixture Model (GMM) also Multiple Kernel Learning with Augmented Features (AFMKL) is applied to learn an adapted classifier based on multiple kernels and to pre-learn classifiers of other action classes. Limin *et al.*[26] experimented with mid-level and spatio temporal part for action recognition. Motionlets comprises of extraction of 3D regions having high motion saliency and then clustering the candidates from each region followed by the selection of region by using greedy approach having effective candidates. Further representation of mid-level video by using midlevel activation vector. Jingen *et al.*[27] proposed the method for recognizing actions from unconstrained wild videos by capturing both static and motion features. Static features are mined using page rank methodology, while motion features are extracted using motion statistics. Finally, heterogeneous features are combined using AdaBoost for classification. Dong *et al.*[28] proposed hybrid descriptor combing static data from histogram of oriented gradient (HOG) descriptor as well as motion information from motion boundary histogram (MBH). STIP generated using dense trajectory sampling are used by Vector of locally aggregated (VLAD) descriptor to encode the generated information and later on recognition of actions take place. Huimin *et al.*[29] proposed approach based on combining shape and motion information. Detecting of objects from frames using background subtraction than extracting local features such as shape and motion information also global features contour coding of motion energy image. Later recognized activities are classified using multi class SVM classifier consisting of binary hierarchical architecture. Orit *et al.*[30] proposed the approach focusing on local change in motion directions. Pattern based approach is used to capture motion flow of present and next motion element. Also bag of words approach is used to get information from decoupling of the image edges from motion edges.

2.2.1.3. Other Approaches

This category deals with human activity recognition approaches based on LBP, soft computing etc. Some of prominent literature following these approaches are discussed below.

Chen *et al.*[31] experimented Depth motion maps with local binary patterns action and introduced a recognition framework using kernel-based extreme learning machine

(KELM). Feature level fusion having LBP features from three different depth motion maps i.e. front, side and top are merged which are further used to classify the actions at decision level fusion using KELM. Two experiments are performed one having two out of three splits is used for training and other experiment consisting of one out of two splits is used for training. Ahsan *et al.*[32] experimented directional motion history image (DMHI) with LBP. DMHI consisting of left MHI, right MHI, down MHI, up MHI is used to represent the spatio temporal features in single image which is further used by Local Binary pattern to extract features for making histogram of feature vector. Along with LBP histogram, shape feature of the action represented as a histogram of selective silhouettes are also used. After the training SVM is used for classification purpose. Well comparative study on different variants such as histogram of LBP from MHI, MHI+ Silhouette image, DMHI and DMHI+ Silhouette image show that DMHI + Silhouette image performs state-of-the-art. Silhouette based Human Action Recognition only focus on singe view but Alok *et al.*[33] proposed the method based on multiple view combing contour-based posture structures from silhouettes and unchanging rotation local binary patterns. Methodology involves three steps first comprises of detection of peoples my means of background subtraction, second involves collective rule invariant contour-based pose structures from silhouettes then even rotation invariant local binary patterns (LBP) are extracted, and finally SVM Classifier is used for classification of different activities performed by people. Vili *et al.*[34] presented two novel texture methods first one uses templates to capture the movements of human action and then describes it with texture features by using Local Binary Pattern. Also, dynamic texture descriptors are extracted from the spatio temporal space by using LBP-TOP having LBP description from three orthogonal planes. Thanh *et al.*[35], proposed a method based on revisiting of local binary pattern based texture models. Local Binary pattern contains the motion points of the surroundings, and self-similarity operator is used to highlight the traced shape from unconstrained videos for action recognition. Acton recognition is done well when using the feature vector. Swarup *et al.*[36] proposed method for binary silhouette generation using directive local binary pattern containing orientation and intensity variation features. Further, these features are combined with Edge Orientation Histogram for a more advanced feature and at last SVM is used both for training and classification purposes. Enqing *et al.*[37] experimented the process for generation of the local binary pattern by combining Static History Image and Motion History Image, making it more capable for containing more motion information along with principal component analysis is used as a descriptor.

Mohammad Helmi *et al.*[38] proposed method based on triaxial accelerometer parameters such as amplification, the correlation between axis and standard deviation, which is input to Fuzzy Inference System (FIS), and deployed to classify four variants of basic activities such as going upstairs and downstairs, moving forward and jumping. Jyh-Yeong *et al.*[39] also experimented Fuzzy rule based system with template matching approach. Object is extracted by foreground removal and is converted to binary image further template matching approach is applied, and posture sequence is classified using fuzzy rule-based system. Manabu *et al.*[40] proposed approach based on MEMS (Multiple micro electro mechanical system) which stores the numerical data from sensors of numerous daily life activities. Rules are generated by using the daily life actions and later on the mapping of actions with the data based on fuzzy rules is done to classify the actions. Recently, Prudhvi *et al.*[41] proposed intelligent fuzzy approach based on inertial measurements units and smart shoes for activity recognition. Four inertial measurement units are mounted on bilateral thighs and shanks of human being and ground contacted forces are measured using extended Kalman filter which is capable of recognition of six basic activities such as going upstairs and downstairs, moving forward, backward, jogging, walking and standing.

Vezzani *et al.*[42] resented an interesting approach of online action recognition of actions by means of HMM batteries taking into consideration all the possible time boundaries and action classes. A suitable Bayesian normalization is applied by repeating matching of the instance and the templates over sliding, overlapping temporal windows making this method suitable for real time scenario. Jingen *et al.* [43] drafted high level approach for action recognition using attributes. Manually driven attributes extracted according to intraclass variability are combined with data driven attributes which are automatically inferred from training data to make attributes set more descriptive. Latent SVM approach is used to learn the latent attributes to make the proper ranking of each attribute. Hamed *et al.*[44] described an approach for the learning of actions using hierarchal grammars. Grammar patterns capture hierarchical temporal structures concluded from finite state machine and are trained using latent structural SVM while sub-actions are learned automatically. Alok *et al.*[45] proposed the method in which objects are classified using template based approach and their silhouettes are generated after that rule based approach is used to classify the actions recognized. The method is very much capable of recognizing seven types of primitive actions with higher accuracy also suitable for real time environment systems. Saif *et al.*[46] proposed method combining ultrasonic sensor and rule based approach for classifying static as well as dynamic activities. Hexamite Hx19 sensors location measurements are used to make rules which are generated using information of adjoining time steps of finite state automata machine. Sreemananath *et al.*[47] presented a high level approach of Action Bank in which a large set of action detectors at various viewpoints and scales are stored. The semantic transfer is there from input video and is combined with SVM classifiers for more descriptive recognition. Anali *et al.*[48] proposed three phase model in which first step is video decomposition, second is video description and last is video classification. A sparse set of meaningful key sequences are used to represent videos and to study inter class and intra class similarities. Afterwards these are projected with bank dictionary by using Inter Temporal Act descriptor (ITRA) covering the relations between different key sequences. Table 2.1 below shows a summary of various conventional machine learning based approaches for human activity recognition.

Table 2.1 Summary of Handcrafted Feature Based Human Activity Recognition Approaches

Sr. No	Paper reported	Description	Pros	Cons
1	S. Samanta *et al.,* [7] (2014)	STIP features are detected and their description is through facet model. 3D HAAR wavelet is used by bag of words later on SVM is used as classifier	Easy to implement and cost efficient without any extra hardware	Method outperforms in case of large frames while calculating STIP from videos
2	K. Mozafari *et al.,* [8] (2011)	Histogram of optical flow (HOF) and Harris detector algorithm is used for feature extraction along with it least square Twin SVM used for classification	Usage of LS-TSVM makes the classification process four time faster	Performance increase with fold rather an evolutionally algorithm with flexible penalty should be used.
3	M. S. Bella *et al.,* [9] (2014)	STIP are extracted using Harris operator in time dimension and clustered and bag of features are built further adaboost SVM is used as classifier	Method is very effective and robust by using Adaboost classifier	Activities like Walking is not classified properly and is able to classify only few actions
4	S. Ma *et al.,* [10] (2015)	Space time segments are used to build spatio temporal trees from training data later on action words are generated for each action	Trees are further re used for different action sets without any training	Trees ignore the body parts which are not discriminative in nature
5	E. Vig *et al.,* [13] (2012)	Only limited number of descriptors are sufficient for recognition purpose in saliency detection along with-it eye movements are	Method works well for professionally edited videos	Coarse scene gist and a detailed foveal view of a scene is not properly used

		used to explore saliency guided descriptor pruning		
6	C. Wang *et al.*, [14] (2013)	Pose based approach containing information by using data mining rules to get spatio temporal structures from action representation	Method is interpretable because only information of 14 joints is needed, compact and computationally efficient.	Method is not able to recover in 3D and also there is problem of occlusion
7	H. Wang *et al.*, [15] (2011)	Optical flow fields on videos are used to extract dense trajectories and bag of features approach is used to evaluate the performance of dense trajectories	Problem of moving camera is resolved by this method as it focuses on foreground motion	Activities like handshake, ball shoot are not properly recognized by this methodology
8	X. Peng *et al.*, [16] (2013)	Refined dense trajectory-based DT-MB captures binary motion boundary from optical flow and regions without motion boundaries are deleted and central points of remaining are patches are averaged.	Method optimizes computation and memory without any loss in performance	Only improvement of 0.2% in the results for KTH dataset by combining all modalities
9	Y. G. Jiang *et al.,* [17] (2012)	Dense trajectories are fused with local and global motion reference points making this method suitable for recognition of actions in which there is motion between the objects	Method is capable of recognizing of actions without background subtraction	Clusters are recognized for identifying trajectories which make it difficult to recognize containing dense trajectory
10	M. Jain *et al.,* [18] (2013)	Visual motion is decomposed into space time trajectories further used by DCS descriptor for description purposes Also, VLAD coding methodology is used for recognition	Methodology is simple and easy to implement for extraction of trajectories	No external vision cues are used which leads to no improvement in the performance
11	B. Fernando *et al.*, [19] (2015)	Features from the video is used by Ranking machine which orders them according to temporal information of video and SVM is used as classifier	Method easy to implement and gives accurate result with fast computations	VideoDarwin method is not capable of performing better with fisher vectors
12	L. Wang *et al.,* [20] (2007)	Foreground detection of input video leads to human silhouette extraction which	Feature extraction is very simple and easy to extract also	Only one cue is used instead multiple cues

		is used by Kernel Principal Component and Factorial Conditional Random Field for projection trajectory	method is not dependent on features used	should be fused for getting better results
13	H. Zhao *et al.,* [21] (2009)	Human postures from video are calculated further silhouette are calculated and Radon transform is done for posture recognition using edit distance database	Robust to Noise as it captures boundary as well as internal content of video	Using of Silhouette with shadow makes Radon transform not well in recognition
14	N. Ikizler *et al.,* [22] (2009)	Grammars using Finite state machines are used for temporal structures and SVM for classification	Parser able to count action instances correctly and also sub actions	Detection of Labelled frames is not to the point
15	F. Niu *et al.,*) [23] (2004	Shape and Motion features are combined for human activity recognition considering all views and Hidden Markov Model is used for activity representation	Usage of all viewing angles make the methodology view invariance and robust	Not able to classify complex activities consisting of two or more actors
16	M. Chen *et al.,* [24] *(*2015)	Spatio temporal features extracted are used by Lie Algebraized Gaussians and Gaussian Mixture Model are trained also SVM is used as classifier	Usage of Gaussian mixture model doesn't make each feature specific to one visual word	Realistic data is not possible to be processed using this method
17	L. Jingen *et al.,* [27] (2009)	Framework for recognizing real actions in wild environment using adaboost C.45 as classifier. Static and motion features extracted goes for pruning and vocabulary learning	This method explores extraction of actions from unconstrained videos making it extra ordinary	Not able to classify actions in which there are multiple motion features
18	D. Xing *et al.*, [28] (2014)	STIP extraction is done from video along with hybrid feature descriptor consisting of static and motion information later on vector of locally aggregated descriptor is used as video encoder	Improved recognition rate because of combination of HOG and MBH leading to hybrid descriptor	Usage of PCA makes the results non-stable and controllable hence it is not merged
19	H. Qian *et al.,* [29] (2010)	Automatically able to detect classes from the videos and their relation with the actions for recognition	Dual framework able to recognize actions and scene in the video	Not worked well for classes having dependence on scenes

20	C. Chen *et al.,* [31] (2015)	LBP along with depth motion maps from three views i.e. front, side and top are fused at feature level and decision level. Kernel extreme learning machine is used as classifier	Method is easy to be implemented in real time as it is having processing rate of 30 frames per second	Similar actions like hand catch and hand throwing are not well classified
21	S. M. M. Ahsan *et al.* [32] (2014)	Directive Local Binary Pattern along with Motion History Image containing spatio temporal features are used for Histogram later on SVM is used as classifier	Method is able to classify different actions with high recognition rates	Not suitable for change in view point or any scaling of point or rotation of the point
22	A. K. S. Kushwaha *et al.,* [33] (2017)	Mutual scale invariant contour-based posture structures from silhouettes and unchanging rotation invariant local binary patterns (LBP) are mined, after that SVM is used as classifier	Method requires less computation resources for processing	Bending and sitting activities are not properly recognized
23	V. Kellokumpu *et al.*, [34] (2011)	Temporal templates are used to capture the motion information later on dynamic texture descriptor is used to describe the captured human movement and SVM is used as classifier	Can be used in online recognition and in real time classification	Jogging and running activities lead to most of the misclassifications during recognition
24	T. P. Nguyen *et al.*, [35] (2013)	Trajectories are used to represent the spatial information of moving parts of tracked body along with LBP is used and SVM is used as classifier	Able to recognize actions in unconstrained environment having complex background	Moving background is still a problem in classification process
25	S. K. Dhar *et al.,* [36] (2016)	Directive Local Binary incorporates binary silhouette images along with edge-oriented histogram to extract features which are used by SVM for training and classification purpose	Method is capable of recognizing sliding and running activities very effectively	Not able to recognize much activities having multi-person interaction and partial occlusion
26	E. Chen *et al.,* [37] (2017)	Motion cues are determined from Depth motion maps of three projection views and local binary pattern is used for feature representation	Usage of Extreme kernel-based machine learning makes the technique more reliable	Similar actions are not classified properly

		along with two type of fusions i.e. feature and decision		
27	P. T. Chinimilli *et al.*, [41] (2017)	Smart shoes and inertial measurements units with fuzzy inference algorithm which detects the activities based on ground contact forces.	Extended Kalman filter is used which detects the transition between the activities very smoothly	Only possible to detect the limited number of activities
28	J. Liu *et al.*, [43] (2011)	High level semantics attributes are learned directly from the training data as well as manually data is also used and K-Nearest Neighbor is used as classifier	Data driven problem is optimized in selecting attributes from the most discriminative ones	Critical actions are difficult to implement in the form of attributes
29	A. K. Singh et al. [45]	Silhouettes generated for different objects and are classified by template matching approach at last rule-based classifier is used for activities classification	Fast computations and high accuracy make the method to fit in real time	Suitable only for classification of limited number of activities
30	S. Okour *et al.,* [46] (2015)	Rules are generated by measurement of distance from fixed Ultrasonic sensor for particular activity by using finite state machine in health smart home	Method is robust, easy to deploy as well as is inexpensive in term of hardware	Not used in Ambient Assisted Living because of bulky wearable devices
31	S. Sadanand *et al.,* [47] (2012)	High level approach consisting of action bank for various small actions pooled together from video labels. Semantic transfer from Action Bank to CNN is done for classification purpose	No need of any feature extractor directly classification sing high level representation	Method is not capable for classification in streaming videos and in real time applications

2.2.2. Deep Learning Based Approaches

Deep learning is a specific subset of Machine Learning, which is a specific subset of Artificial Intelligence (Fig. 2.4). Artificial Intelligence is the broad mandate of creating machines that can think intelligently. Machine Learning *is one way of doing that*, by using algorithms to glean insights from data. Deep Learning *is one way of doing that*, using a specific algorithm called a Neural Network.

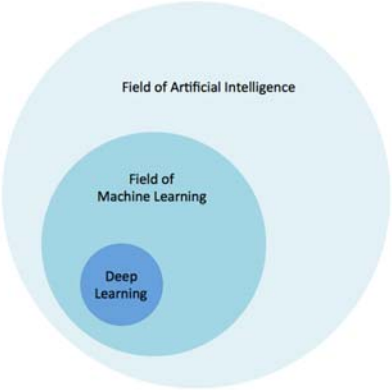

Figure 2.4 Relationship among AI, ML and DL

The Deep Neural Network architectures address activity recognition problem by applying multiple levels of non-linear operations. Deep artificial neural networks contain multiple hidden layers as opposed to shallow network that has just one so-called hidden layer. Multiple hidden layers allow deep networks to learn features of the data hierarchically, because simple features learnt by lower layers recombine from one layer to the next forming more complex features at top layers.

Deep learning methods include automated learning of features using CNNs and RNN's. Deep Learning methods are nowadays at great peak. These methods are further classified into supervised learning and non- supervised learning. Joe *et al.*[49] presented a method in which spatio temporal features from long term video frames are learned over LSTM using GoogleNet and AlexNet Models. Along with them RGB and Optical flow are used as the source of input for recognition purposes, and feature pooling is used to classify actions. After the huge success of CNN;s over images these are extended for 1 million videos Sports 1M dataset [50] having huge variation of activity classes approximately 487. In addition, 2 spatial resolutions - low resolution context streams and high-resolution fovea streams are used to attain good results and lessen training time. To validate the results of other interesting data sets, the perception of transfer learning was used for data set UCF 101 and slow fusion networks, and better outcomes were achieved in only a few layers of fine-tuning, rather than from scratch learning. Fuqiang et al. [55] projected a technique which is based on deep learning to identify sports activities. To do this, use multiple smartphone sensors such as gyroscopes, accelerometers, and magnetometer data. Stacked denoising auto encoders are used to eliminate training expertise. The learning activity model from live video is still an interesting job, but [56] proposes the use of deep networks and active learning. Dual types of preliminary and incremental learning are presented, and in the preliminary phase, there are very few marked and unlabeled occurrences for automatic encoders to learn. At this stage, the model based on activity recognition from the above stages and active learning are used. The main feature set of the active class is reduced in a tab-free method. Convolutional network towers for space level and time level are merged at the softmax level, but Christoph et al. [57] Experiments at the final convolutional layer and class calculation layer fusion layer, and achieved improved outcomes. The space for 2-Dimensional features and time fusion is used for two pre-trained ImageNet models. Moez *et al.*[51] projected sequential learning without using any prior knowledge by extending CNNs to 3D, which enables it to learn spatio temporal features automatically. Moreover, Recurrent neural network having Long Short-term memory (LSTM) is trained to classify the learned features having a special type of node Constant Error Carousel (CEC) which makes it constant error signal propagation through time. Shuiwang *et al.*[52] presented 3-Dimensional Convolutional Neural Network architecture consisting of single hardwired layer, 3 convolution layer, 2 subsampling layer, and 1 fully connected layer. Supervised deep architecture generates multiple channels of information from adjacent input frames and performs convolution and subsampling separately in each channel. Scores are calculated by using information from all the channels. Bharat Singh *et al.* [53] proposed fine grained act detection method based on multi stream bi directional RNN. Spatial stream consisting of motion and appearance information for full frame and person centric frame is used to analyze short chunk of videos. Fabian *et al.* [54] proposed largescale newer dataset known as ActivityNet consisting of 203 activity classes consisting of 137 untrimmed videos. ActivityNet dataset is compared with other popular datasets, and their applications are also discussed such as trimmed classification, untrimmed classification and activity detection in videos. Recently, Madhuri *et al.*[55] proposed deep learning based approach based upon wrist worn accelerometer sensor of four different subjects. After preprocessing the CNN classifier is used to extract features and classify basic forearm movements. Deep Learning methodology directly assigns the label to the video but Wangjiang *et al.*[56] proposed the key volume mining deep framework which assigns the labels in backward pass and mines key volume for each class in forward pass. Proposed network is optimized and follows unsupervised learning. Hakan *et al.*[57] proposed the method of representing motion information of whole video in the single image known as dynamic image. Dynamic images and Average rank pooling results into four stream architecture for classification of actions in videos. RGB frames, Dynamic Image, Optical flow stream and Dynamic optical flow stream are used for score averaging from which class prediction is made. The learning activity model from a live video is still an interesting task. [58] proposes the use of deep networks and active learning. Two different types of learning, namely initial and incremental, are presented, and in the initial phase, there are very few marked and unlabeled instances for automatic encoders to learn. After that model learned from above staged is merged with active learning. Mainly this decreases the finest set of features for action class in an unlabeled manner. Zelun *et al.* [59] A method based on the RGB D modality, which is proposed for generating a sequence of atomic 3 D streams. Later, the RNN is used to predict the classification actions from these 3D streams. This model is suitable for any input modality such as RGB, RGB-D and depth and captures long-term motion dependencies and spatial time relationships. Juarez *et al.* [60] proposed approach for activity detection in the indoor environment by using a single static camera. CNN architectures are grouped into two classes one containing pre-trained models and other containing fully trained models. Pre-trained group contain AlexNet, GoogLeNet and SqueezeNet models trained on ImageNet dataset also SVM is used as a classifier. Whereas, a fully trained group contain GoogLeNet trained from scratch on KitchenSet dataset also RNN is used as a classifier. Scores from both are used for prediction of action. Eunbyung *et al.* [61] fused spatial features with temporal features using two methods. Optical flow is combined with last convolutional layer of deep architecture model. Secondly, varying spatially multiplicative last layers of different CNN's trained on different subjects are combined. Jun *et al.* [62] experimented Hidden Markov Model with deep CNN's. The automatic ability of learning the features from raw data is implemented by CNN whereas HMM is used for modelling time sequence along with training of CNN-HMM model with Viterbi algorithm. Tushar *et al.* [63] proposed method in which Gaussian Mixture Model is used to make binary frames by removing background. Later on, binary frames are used to calculate binary motion image (BMI) which is used as input to CNN for training and testing purpose. CNN are not only used with hand crafted features rather Earnest *et al.* [64] used CNN with fuzzy inputs from motion capture information. Membership functions are derived by using the distance of ground from left hand, right hand and pelvis which is used for feature representation in action recognition and is used to train CNN. Usually large amount of data is needed for training of deep networks and to perform in consistent manner. But, Sheng *et al.* [65] proposed two stream fully convolution networks which can be used by using limited parameters keeping the performance consistent. Appearance features and motion features are fused along with Temporal pyramid pooling is used as pooling layer for action recognition. Vivek *et al.* [66] proposed the extended version of LSTM known as differential Recurrent Neural Network capable of automatic learning of dynamic saliency of spatio temporal features from actions by using the backpropagation algorithm. Lin *et al.* [67] handled the problem of learning of 3D convolution kernels by factorized spatio temporal convolutional networks. Lower layers i.e. Spatial layer is used to learn 2D convolutional kernels whereas upper layers i.e. temporal, is used to learn 1D convolutional kernels without any additional training video. Guilhem *et al.*[68] combined motion and appearance information with CNN to form Pose based CNN (P-CNN) network descriptor. Time to time joint information in the form of optical flow along with RGB frames are fused together for score calculation and are later used for action recognition. Lin *et al.*[69] proposed method based on combing Slow feature Analysis with deep learning methodology. Two layered Slow Feature Analysis with 3D convolutional layer and max pooling layers are used to learn hierarchal structure which is generic and is fully automated for action recognition [70]. Pushpajit *et al.*[71] combined different vision cues such as RGB data, skeletal data and depth data from RGB-D sensor. Different input modalities containing images are trained separately using convolutional model, and their fusion score is combined at decision level for action recognition. Recently inspired by the success of two stream neural network Zhigang *et al.*[72] proposed multi stream CNN model which consists of three two stream networks having motion based streams and appearance based streams. Improved selection of human regions is done in two phases R1 and R2. R1 consists of full body area and is extracted using Improved Block sparse Robust Principal Component Analysis Whereas R2 is extracted using motion saliency area of the action. Depth and RGB modalities are fused later but Pichao *et al.*[73] proposed method based on Scene flow vector, which consists of RGB and optical as single modality. Scene flow to Action map methodology consists of scene flow vectors having long spatio temporal information for action recognition. Guha *et al.*[74] presented an approach in which 2D CNN are extended to 3D CNN yielding a fully automated deep learning based framework having RNN as classifier for each learned spatio temporal sequence in each time step. Yongmou *et al.*[75] presented the strategy of extracting hand crafted features by using gyroscope and accelerometer sensor data. Unsupervised feature learning by using PCA, sparse auto encoder makes the framework to learn features automatically from massive training data. Ryoo *et al.*[76] compared the available methods of learning convolutional neural networks such as pooling, image based CNN and RNN's on robot centric videos having robot human interactions or the videos in the outdoor regions. Basura *et al.*[77] proposed a rank pooling machine in which Temporal ordering of the video is preserved by training linear ranking machine according to ranking of the frames capturing appearance and evolution over time by supervised learning. Earnest e*t al.*[78] presented the approach in which action banks extracted in two ways are used in convolutional neural network as input. Action banks are used in extracting the feature and are used by convolutional neural network. Recently Cheng et al.[79] presented the approach in which Body Activity Recognition using data from wearable sensors. Classifying of the actions is done by using different approaches such Machine learning algorithms artificial neural network. is Some of the recent works are contributed by [80] in which dynamic images are constructed using Dynamic Depth, Dynamic Depth Normal and Dynamic Depth Motion Normal. Then these dynamic images are used for further training and classification purposes.

Over the past decade, numerous works have been done on the recognition of human activities. In literature, different kind of methods exists for activity recognition. After the introduction of Deep Learning models, researchers have proposed several frameworks with significant improvements. N. Jaouedi et al. [81] proposed a method using GMM and Kalman filter for motion tracking and GRU(Gated recurrent unit) for activity classification. Li et al.[82] Obtained motion feature, frequency feature and statistic feature by pre-processing of the raw input data and used these feature for training the CNN model. Jayabalan et al. [83], took advantage of less dimensionality of join data and proposed a CNN model for training and classification. Feichtenhofer et al.[84] presented a two-stream model using video frames and optical flow. A multi-stream model using video frames and dense optical flow has been proposed by Simonyan and Zisserman [85]. Ijjina et al. proposed a method using a genetic algorithm for analyzing CNN classifiers weights [86]. Varol et al. [87], has introduced a method based on spatial and spatio-temporal features using RGB values and flow values as training input of the model. N. Nair et al. [88], presented a model using Temporal Convolutional Network(TCN) for action recognition using wearable sensors. Mo et al. [89], proposed a skeleton data based approach. A CNN model with 6 hidden layers used for feature extraction and classification was done using multilayer perceptron. Karpathy et al. [50] proposed a slow fusion based approach using sequential video frames for training, thus improving temporal awareness of the network. In [90] Ronao et al. proposed a multilayer CNN model using single dimension raw time series sensor data. In [91], the authors presented a depth map based fusion model using weighted hierarchical depth motion maps (WHDMM). Parka et al. [92] presented a Joint angle feature from human silhouette based approach. Alberto monte et al [93] proposed a pipeline architecture using 3D CNN for temporal activity detection by splitting video into 16 frame clips, RNN is used for prediction of class probability, the maximum average class probability is considered as predicted class. In [94], Liu et al. presented a model using depth sequence and joint vector for network training which is view invariant and time invariant, however, consumes a lot of pre-processing time. Dobhal et al [63] used Binary motion Images for training the CNN model and Gaussian Mixture Model (GMM) for background subtraction. In table-1 we have presented a summary of these state of art methods using deep learning approach in terms of their reference, short description, model for feature extraction and classification, datasets and results on those datasets and finally remarks. Table 2.2 presents a summary of some human activity recognition approaches using deep learning based approaches.

Table 2.2 Summary of Deep Learning Based Human Activity Recognition Approaches

Reference	**Input**	**Description**	**Feature extraction**	**Classifier**	**Dataset/ Results**	**Remarks**
Neziha Jaouedi *et al.* [81] (2019)	Video frames	Used a hybrid approach for human activity recognition based in motion tracking and classification.	GMM and Kalman Filter	GRU	KTH/ 96.30 UCP Sports/ 89.01 UCF 101/ 89.30	Gives good result for controlled environment dataset but for complex dataset, it can be further improved.
Li *et al.* [82] , (2017)	Motion feature, Frequency feature, Statistic feature	Raw spatial data has been pre-processed in motion, frequency and statistic feature and fed as input to the CNN for training. At output layer, softmax classifier has been used for classification.	CNN	Softmax	Own dataset prepared in lab/ 86.7	Produces better result lower size kernel. For further studies, the model can be trained on larger datasets
Jayabalan *et al.* [83], (2017)	Feature vector	HAR Model using joint data, taking advantages of less dimensionality of joint data features. Making the model less complex and more faster than RNN.	CNN	CNN	CAD-60/ 87.0	We can obtain Information regarding joint data from many source like kinetic sensors, lidar, stereo camera or smart clothing, making the model robust. Other CNN model may also be

						integrated with this model to classify unrecognized activities.
Feichtenhofer *et al.* [84] (2016)	Video frames + Optical flow	Two-stream approach strengthened by maintaining correspondence between spatial and temporal features i.e. it is useful to know which parts of the image are moving and "what" is depicted in the moving parts of a video by using fusion within a convolution layer to combine the separate streams.	CNN	CNN	UCF-101/93.5 (with IDT features) HMDB-51/69.2 (with IDT features)	Establishes learning correspondence on both spatial and temporal convnet features.
Simonyan and Zisserman [85]. (2014)	Video frames, dense optical flow	Multi-stream model using spatial stream for action in formation in the still frame and motion information from the dense optical flow. Output Scores of both the streams is fused to find class of the activity.	Two stream CNN	Softmax, SVM	UCF-101/88.0 HMDB-51/59.4/91.56	By Separating the two streams, model exploits the available huge amount of annotated image data. Camera motion has to be handled explicitly in this model, which can be improved further.
Ijjina *et al.* [86], (2016)	Action bank features	A genetic algorithm has been used for Initializing the weights of CNN classifier making the model less prone to classification errors.	CNN	NN+ELM	UCF-50/99.98	No problem of overfitting nor stuck to local minima. Further, model can be extended for spatio-temporal feature learning.
Varol *et al.* [87], (2016)	RGB values, Flow values	Improved action recognition accuracy by using a space time CNN over a large	$LTC_{flow+RGB}$	IDT	UCF101/92.7 HMDB51/67.2	With limited training data, we get improved performance.

		temporal convolution having large temporal depth at a cost of low spatial resolution. Signifies the use of high level optical flow.				Result is highly affected by flow quality.
N. Nair [88] (2018)	Sensor data	Proposed an encoder decoder based technique using TCN (Temporal Convolutional Network for Activity classification.	TCN	TCN	UCI HAR(ED-TCN)/ 94.6 UCI HAR(Dilated-TCN)/ 93.8	Proposed method may be further improved for recognition of more fine grain action classes.
Luo *et al.* [59], (2017)	Depth information	An Encoder-Decoder based architecture using Recurrent Neural Network. CNN Encoder learns motion features covering long-term dependencies. Architecture complexity reduced by using RGB-D modality.	CNN+LSTM	RNN	MSRDailyActivity3D/86.9 (Depth based) UCF-101/79.3 (RGB Based)	Dense trajectory representation can further improve performance by reducing background motions.
Mo *et al.* [89] (2016)	Skeleton data	For input to skeleton data has been used. Feature extraction has been done using CNN model with 6 hidden layers and classification done using multilayer perceptron.	CNN	Multilayer Perceptron	CAD 60/81.8	Effort at data processing and feature extraction level reduced. More feature can be added to improve the model. Also Multilayer perceptron can replaced with RNN
Karpathy *et al.* [50] (2014)	Video frames	Each CNN is trained on many consecutive parts of the video and performed slow fusion to increase temporal learning of the network.	CNN	CNN	Sports-1M/63.9 (Video Hit@1) UCF-101/66.0(mAP	Fusion improves temporal awareness of the network.

					fine-tune top)	
Ronao *et al.* [90] (2016)	1D sensor data	Single Dimension Time series raw sensor data used for training multilayer CNN with alternate convolutional and pooling layers. Finally fully connected layer has been used for classification.	CNN	CNN	dataset from 30 subjects/ 94.79 (raw sensor data)	Comparable performance with state of art models. Can be integrated with SVM.
Wang *et al.* [91], (2015)	Depth maps	Fusion based approach using weighted Hierarchical Depth Motion Map. Convnet is trained on WHDMMs generated from top front and side and result is fused for final classification.	CNN	3CNN	MSRAction3D/ 94.92 MSRAction3DExt/94.35 UTKinect-Action/ 92.93 MSRDailyActivity3D/80.63 Combined/91.56	Convnet trained on a large dataset gives better performance, even in changing view conditions. Suitable for less number of classes.
Parka *et al.* [92] (2016)	Human Silhouette	Joint angle features (28 joint angles corresponding to 14 body parts) were used to train RNN with spatio-temporal feature matrix. These features were extracted from human silhouette.	Handcrafted	RNN	MSRC-12/99.55	Use of time sequential encoding produces better results compared to HMM and DBN based approaches.
Albertomonte *et al.* [93], (2016)	Video frames	Pipeline architecture for temporal activity detection that splits video into 16-frame clips, each being fed to C3D for feature extraction which in turn acts as input to RNN for	CNN	RNN	ActivityNet/59.38	Detect the segment of video containing activity, hence able to find temporal extent of he activity. To further improve end to

		predicting class probabilities. Class probabilities are then averaged over all clips and maximum probability class is considered as predicted class				end training model using 3D-CNN and RNN can be used.
Liu *et al.* [94], (2016)	Depth sequences, Joint Vector	3D deep CNN used to learn spatiotemporal features directly from raw depth sequence and joint vector for each sequence. Learn features are fused and classified using svm.	CNN	SVM	UTKinect-Action/96 MSRAction3D/84.7 (Cross Subject Test)	Time invariant and view invariant feature representation. To make the model insensitive to different objects a lot of time is consumed in pre-processing to obtain depth sequence.
Dobhal *et al.* [63] (2015)	BMI	Binary Motion Images (BMI) are generated by combining a sequence of image into a single image. Foreground is calculated using GMM. BMI is used for training of the CNN for Classification.	CNN	CNN	MSRAction3D/97.5	Model does not have effect of speed of action, partial occlusion and holes. Further it can be extended to train models using 3D Depth Map. 2-D Image maps small movements within silhouette cannot be detected.

2.2.2.1. Convolutional Neural Networks

Deep network architecture popularly used for activity feature extraction and classification is Convolution Neural Network (ConvNet). A simple ConvNet (Fig. 2.5) is a sequence of layers, and every layer of a ConvNet transforms one volume of activations to another through a differentiable function. There are three main types of layers to build ConvNet architectures: Convolutional Layer, Pooling Layer, and Fully-Connected (FC)

Layer (exactly as seen in regular neural networks). These layers are stacked to form a full ConvNet architecture**.**

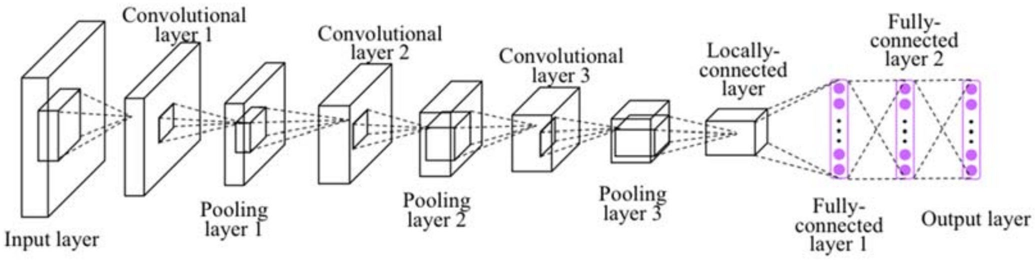

Figure 2.5 Convolutional neural network framework [95]

Convolutional layer computes the output of neurons that are connected to local regions in the input, each computing a dot product between their weights and a small region they are connected to in the input volume. ReLU (rectified linear unit) layer apply an elementwise activation function, such as the max(0, x) thresholding at zero. This helps in reducing the computations by converting all negative value to zero and hence fasten the training. Pooling layer performs the downsampling operation along the given dimensions to reduce the size of the input. Lastly, FC layers compute the class scores to classify the input into corresponding categories. Similar to the ordinary neural network, each neuron in this layer connected to all the numbers in the previous volume.

Convolutional networks have been achieved remarkable results in action recognition and classification task from images. The capability of deep convolution network to learn complex representations from large visual data datasets makes them suitable for video based activity recognition. However, there are two significant factors that influence video based activity recognition with deep convolutional network. First, the length of time dimension which helps to understand the dynamics in activity videos. Second, large volume of training data to obtain optimum accuracy.

2.2.2.2. Recurrent Neural Network

Recurrent Neural Networks are a superset of feed-forward neural networks but they add the concept of recurrent connections. These connections (or recurrent edges) span adjacent time-steps (e.g., a previous time-step), giving the model the concept of time. The conventional connections do not contain cycles in recurrent neural networks. However, recurrent connections can form cycles, including connections back to the original neurons themselves at future time-steps. Recurrent Neural Networks take each vector from a sequence of input vectors and model them one at a time. This allows the network to retain state while modeling each input vector across the window of input vectors. Modeling the time dimension is a hallmark of Recurrent Neural Networks. LSTM networks are the most commonly used variation of Recurrent Neural Networks.

2.3. Research Gaps

Based on the survey performed on Human Activity Recognition approaches we have identified the following research gaps in this field-

- The trajectory of activities from different viewing directions is different and some of the body parts (part of hand, lower part of leg, part of body, etc.) are occluded due to view changes.
- The other common issues include fixed or moving cameras, scenes having moving or clutter backgrounds, changes in light and view-point, variations in scale, starting and ending state, variations in appearance of individuals and cloths of human etc. These issues and situations make the human activity recognition a challenging task.
- Human activities are performed in a real 3D environment, and cameras only capture the 2D projection of the real scene. Therefore, visual analysis of activities

carried out in the image plane is only a projection of the real activities. This projection of the activities depends on the viewpoint and do not contain full information about the performed activities.

- Most of the work on activity recognition are view dependent and deal with recognition from one fixed view. Recognizing human activities from multiple views has been a challenging task for researchers around the globe and needs a lot of improvement.
- Most of the 3D CNNs that have been examined for activity recognition used RGB input to learn video representations, which is inferior to capture more discriminatory information comparative to depth sequences.
- In earlier models, very short video intervals have been examined; for example, most of the researchers worked on 2, 7, 15, 16-frame video clips. While an activity last few tens of seconds.
- Most of the research have used RGB frames, optical flow frames, skeleton image, depth map, dynamic images individually, whereas only a few blend them to boost the performance.

2.4. Literature Survey of Datasets for Human Activity Recognition

The goal of a Human Activity Recognition System (HAR) is identifying actions or activities done by a person or a group of people. There is a wide range of applications for activity recognition in security, surveillance, health monitoring and intuitive interfaces for machines. Automated comprehension of human behaviour is also a means for developing intuitive human computer interface in entertainment and gaming. Due to the extensive research focus on this topic, the field has evolved to the point that large-scale generalizable and accurate measurement of human activity is possible.

Modern research on this subject is almost entirely focused on Machine Learning and Deep Learning due to the enormous advantage in adaptability, accuracy and more recently, greater speeds over traditional approaches [96], [97]. For the last few decades, a number of distinct deep learning approaches have emerged, which innovate on different parts of the activity recognition pipeline. The current 'state of the art' machine learning models tend to be very data-heavy and require an adequately large dataset to train them. To fulfil this need, a large number of diverse datasets have been published. The selection criteria for use of a particular dataset varies according to the needs of the algorithm being trained. In recent years there have been many reviews on the subject of human activity recognition; however, they have been focused on the recent breakthrough algorithms [96]–[100]. Instead, we focus on the datasets themselves, specifically datasets related to visual modalities, and their evolution towards modern datasets.

2.4.1. Evolution of Modern Datasets

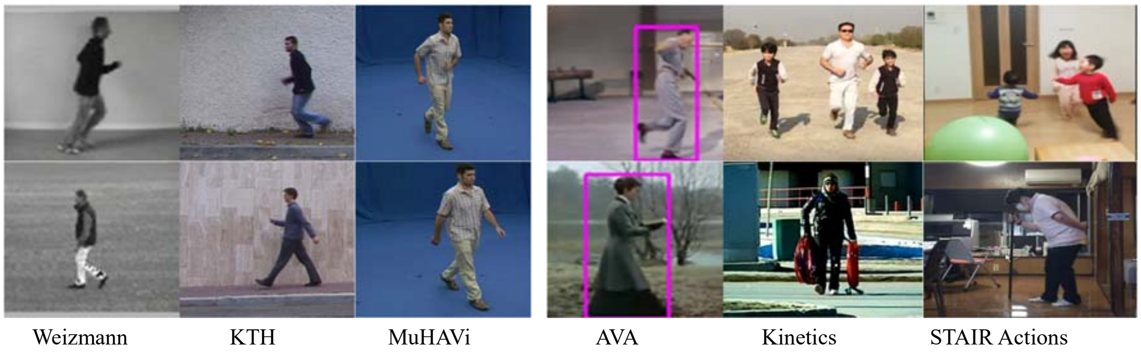

Figure 2.6 Evolution of Datasets from Scripted (left) to Unscripted (right) as seen through common activities: Running (Top Row), Walking (Bottom Row) [101]- [105]

As deep learning systems have evolved over time, becoming more complex and accurate, the popular datasets have also become increasingly detailed over those used for early works [106] (Fig. 2.6). These early datasets, filmed under carefully controlled conditions within labs, included very little variation in the content, e.g. number of actions, number of actors, lighting, occlusion, viewpoints, modalities and size of dataset. Since then there have been many advances in how datasets are created. Better quality datasets allow for more complex models and challenging datasets allow for the evaluation of the robustness and generalizability of these models. It is crucial to include unconstrained action recognition from "real-world" videos as biases in the data eventually impede generality in algorithms, making explicitly controlled environments counterproductive due to overfitting [107]. The implication is that current research requires more general and diverse benchmarks. To this end, datasets are now available with a large variation in the method of creation, source material, quality of video, quantity of labels, and complexity of annotation and generality of content. Ideally, datasets should test the capability of recognition systems for handling contextual cues, partial occlusion, intra-class variability, varying size and pose and contextual cues. They should also provide variation in expression, posture, motion and clothing; perspective effects and camera motion; illumination variation; occlusions and variation in scene surrounding.

With this in mind, we define a modern dataset as one which satisfies the following criteria:

1. Unconstrained Inputs as close to Real World Examples as possible
2. Large Size, preferably high number of classes
3. Exhaustively Annotated

It is necessary, however, to consider that no algorithm can generalize to all known datasets, i.e. there is no 'universal' dataset [98]. In addition, many approaches depend upon inputs which are absent in common datasets which then requires the creation of domain specific datasets. Still, certain datasets have been considered to be benchmark datasets as many general approaches can be applied to them. Moreover, the use of these standardized datasets saves time and allows for comparison of results amongst varied approaches.

2.4.2. Characteristics of Datasets

2.4.2.1. Classes

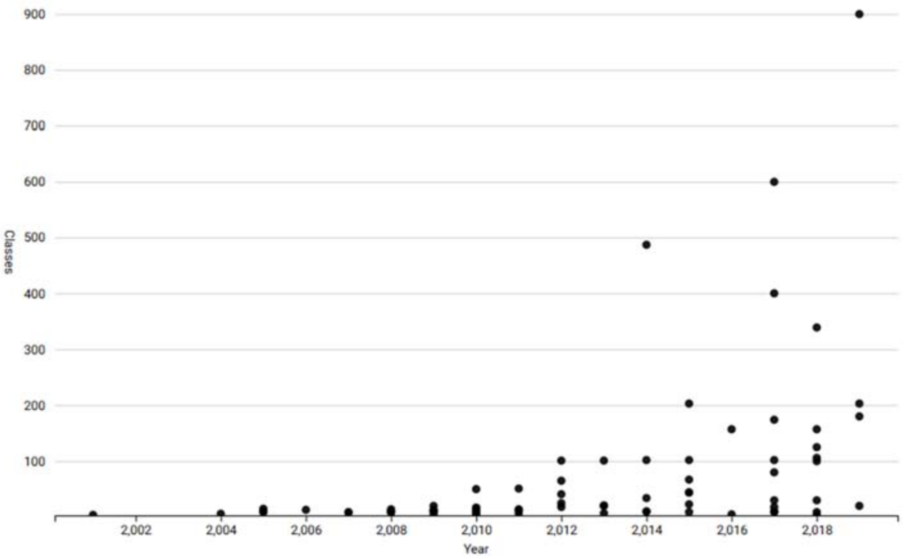

Figure 2.7 The exponential growth in the number of classes in datasets.

The choice of action classes greatly affects the diversity and coverage of the dataset. While certain datasets construct a rich hierarchy of action classes [108], usually the following groupings are present in some form in most datasets [105], [109]–[111]:

1. Person Only: 'low level' solo activities typically involving a single person and usually atomic in nature. The very first research in HAR was focused on activities like *Walking, Running, etc.* [110].
2. Person – Object: 'object manipulation' based activities where the object is essential in defining the activity. Object focused datasets [105], [112], [113] have classes in the form of ('verb + object') such as *Ride Bike, Spray Water, etc.*
3. Person – Person: composite activities involving interactions between a group of people but which cannot be done individually. Datasets such as BEHAVE [114]

allow for the consideration of realistic Person-Person classes such as *Walk Together, Meet, Chase*.

More recently, classes belonging to all these categories are included in larger benchmarks [103], [109], [115] (Fig. 2.7). In contrast, certain datasets forgo the idea of pre-selected activity lists altogether [116] or obtain class lists through text mining [117]. Although the classes considered for HAR have become more complex and realistic, due to inconsistency of class annotations across various datasets, it is difficult to train robust cross-dataset models [118].

2.4.2.2 Focus

A majority of the datasets covered in this survey either consider activities performed during daily life or do not have a very specific domain focus. Besides these, an important sub problem in HAR is sports related activities. UCF Sports [119], [120] , Olympic Sports [121], Sports-1M [50] , Volleyball [122], SoccerNet [123] are all dedicated towards sports activities and they are an important subclass in datasets such as UCF101 [109]. Another application is gaming related actions as in G3D[124], G3Di [125] and MSR Action [126] where classes are chosen for use in hands-free gaming. Kitchen based actions are also a popular dataset choice [116], [127]–[130]. READ dataset [131], [132] describes a recent application in autonomous driving. Other projects such as CAVIAR [133], [134] and ETISEO [135] used actions relevant for surveillance tasks such as such as *Tailgating, Fighting, Shop Entering, and Shop Exiting.* However, modern benchmark datasets typically include all types of classes [103], [108], [109], [115]. For extremely large datasets, classes can be selected using crowdsourcing [110] or dictionary verb lists [105], [115] (Fig. 2.8).

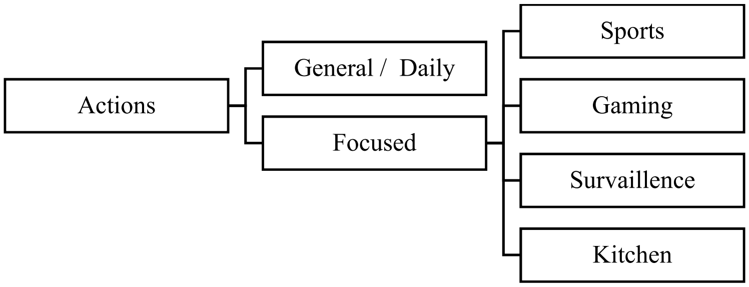

Figure 2.8 Actions and their sub-categories by domain and focus

2.4.2.3. Modality

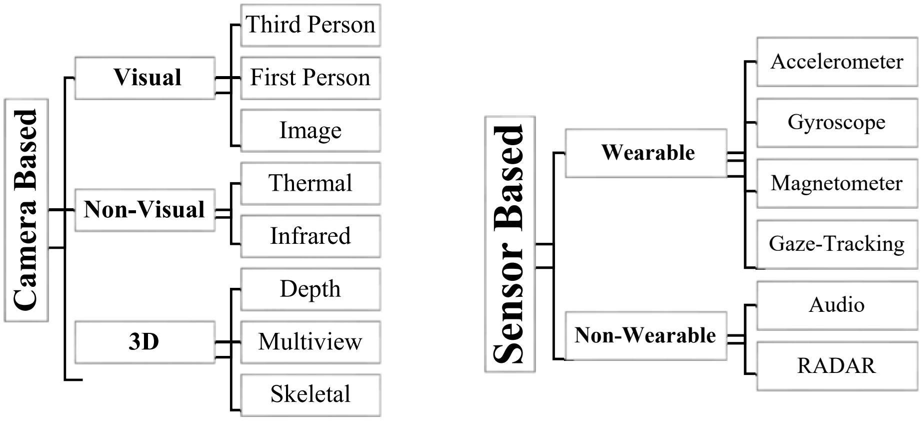

Figure 2.9 Camera and Sensor Based Modality

Human activity recognition, being dependent on both the temporal and spatial context [136], [137], has emerged as a distinct problem from static object recognition or classification(Fig. 2.9). Therefore, HAR datasets usually preserve the temporal dimension, unlike simpler image classification tasks, although there exist large datasets for image based activity classification [112], [113].

The natural representation of datasets is in the form of clips of 2D images, and most datasets use this format extensively. However, after the introduction of low cost 3D sensors like Microsoft Kinect, there has been great interest in using depth information [126]. A detailed description of RGB-D datasets can be found in [138].

Another entirely different class of datasets is recorded using non-visual sensor. This includes several different types of "on-body" (wearable) sensors such as accelerometers and gyroscopes along with ambient and stationary sensors such as audio and RADAR. This class has a great diversity of datasets, such as the Opportunity [139] dataset. Sensor based datasets have been recently reviewed in [140].

2.4.2.4. Data Source

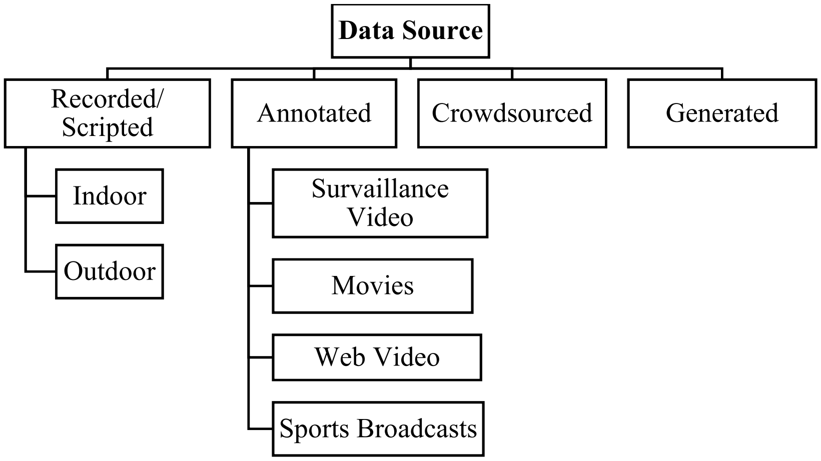

Figure 2.10 Data Source for dataset generation

The source from which dataset is acquired determines to a large degree how well an algorithm trained on it will perform on unseen data. The early datasets consist mainly of video clips containing an individual or a pair of actors performing an activity in an indoor lab setting. These are termed as recorded or scripted datasets. Characteristics of these datasets include: minimal occlusion where actors performing the action are always visible, good consistent illumination, no variation in viewpoints, etc. However, the models trained on recorded datasets tend not to generalize to real world environments which do not have very controlled conditions [100], [141] [142]. Semi-scripted datasets such as [143], [144] ask the actors to perform not the specific actions but give them a task to perform (a recipe) which includes those actions.

In contrast, most modern datasets which include ‘Actions in the Wild’, are collected using one of two main approaches. In the first approach termed as annotated or collected datasets, pre-existing videos are gathered from a variety of sources including movies, sports broadcasts and YouTube. These videos are subsequently filtered and annotated. Web videos (especially YouTube) seem to be the easiest and most widely used source for modern datasets [103], [108], [111], [145]. For such large datasets, the data collection strategy is also an important factor and web search queries have become the standard and easy way of collecting the videos [105], [108], [115]. Annotated datasets are not constrained by controlled conditions and there is a natural variation in viewpoint, lighting and the actors performing the actions. Moreover, with these approaches, it is possible to have datasets with many more categories and a lot of variation within the same category as the videos can be sourced separately. However, this approach effectively limits the dataset to the visual modality only.

The second, “Hollywood in Homes”, approach was introduced by the Charades [146] dataset and later used in many datasets like Something Something [147]. These **crowdsourced datasets** involve using crowdsourced actors to film themselves performing various activity videos from diverse household locations and thus neither annotation nor recording is needed. A similar approach called *‘gamesourcing’* was used in G3Di [125] by using data collected from playing competitive games. Finally, another novel and unique approach was introduced in [148] by creating a procedurally generated dataset by using automatic simulation of action in virtual worlds (Fig. 2.10).

2.4.2.5. Annotation Method

For recorded, crowdsourced and generated datasets, the video label is already known at the time of video creation. On the other hand, for annotated datasets, accurate and precise annotation and subsequent verification of labels is essential for supervised learning schemes. While some large datasets use manual annotation by the researchers themselves, the same can be done much quicker using effective crowdsourcing platforms like Amazon Mechanical Turk (AMT) which has become very popular in dataset creation [149].

Many innovative approaches seek to automate the process of annotation by analysis of secondary information available with the raw video itself. For example, Hollywood [150] uses script time alignment and text classification on the screenplays to automatically label the actions in movies, thereby solving many of the problems with manual annotation. Similar methods are used in [151] for annotation of web videos. SoccerNet [123] used text mining techniques to extract labels from sports commentary.

2.4.2.6. Annotation Type

The type of annotation determines the degree to which an action is localized in time and in space. When temporal localization is not an important concern, as in activity classification problems, the entire sequence or video clip is directly labelled with its corresponding class. This is termed as sequence level annotation and is the case for the majority of datasets due to constraints of more complex annotations. For datasets focusing on detecting activities specifically, frame range or action segment annotation specifies an interval related to a particular class [152].

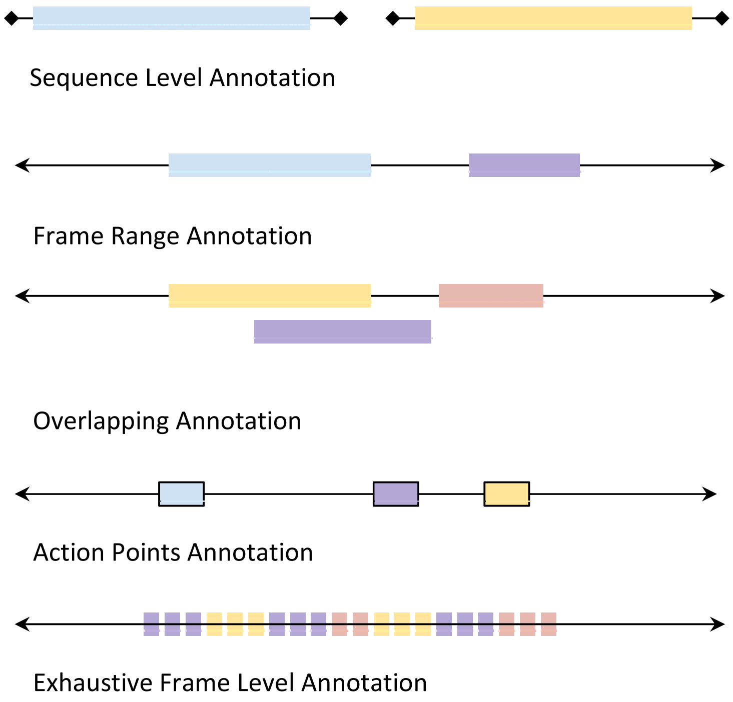

Figure 2.11 Different types of annotations

For activities with well-defined boundaries and having natural window of starting and ending points, this type of annotation is especially relevant, though it can be argued that certain activities do not have such boundaries [123] and that strict boundaries are not as effective as fluid ones [153]. Moreover, such an approach is less useful for online, low latency environments (such as gaming applications) where the action category needs to be known at a predictable point in time [124]. In such instances, action point annotation is used. Action points are the single time instance in which the action can be uniquely identified [154]. A similar concept is the action completion point which is the moment beyond which the action's goal is believed to be completed [155].For spatial localization, datasets also provide supplemental features such as bounding boxes and pose annotations since such a person-focused approach benefits classification [153] (Fig. 2.11).

2.4.3. Video Datasets

In general, an HAR algorithm can have two basic tasks with different evaluation protocols. In action classification task (trimmed activity recognition), the action being performed in a particular clip is identified. HAR thus becomes a multiclass classification problem (with/ without null class). If the sequence level ground truth is available, most datasets use multiclass accuracy as the metric. Although since realistic datasets are unbalanced and long-tailed, some alternative metrics such as precision, recall and F-score are used instead.

Certain datasets also allow for overlapping action labels (e.g.: a person may *Eat* while *Sitting*) both in space and in time. Similarly, the inclusion of a null class in activity recognition in datasets like [156] transforms the problem to a multi label binary classification task in contrast to the forced-choice multi-class task.

In the action prediction task, the period when a particular action is being performed, is predicted, casting the problem into a standard pattern recognition task. For action segment annotation, mean Average Precision (mAP) calculated over the Intersection over Union (IoU) is used as described in [157], [158]. An overview of the metrics and issues in this task can be found in [159].

Less common tasks include spatio-temporal localization which involves predicting the extent of as task in both space (as bounding boxes) and time [103]. Another task is providing localized event captions [160]. ASLAN [161] defines an action similarity labelling task.

Overall, benchmark datasets generally have been found to have a strong built-in bias [152], which can cause research to become constrained. Better evaluation protocols such as cross dataset testing [162] can be used to ensure that the algorithms truly generalize on unseen data.

This section presents the evolution of early and modern datasets for human activity recognition. Table 2.3 presents evolution of early dataset based on their year of evolution, focus, number of activity classes and publicly available link of the dataset. Further, Table 2.3 shows the evolution of modern datasets.

Table 2.3: Evolution of early datasets

Dataset	Year	Focus	Classes	Link
Weizmann Event [101]	2001	Events, Statistical Methods	4	[163]
KTH [102]	2004	Activity Benchmark	6	[164]
Weizmann [165]	2005	Silhouette based methods	10	[166]
CAVIAR [133], [134]	2005	Surveillance	9	[167]
ETISEO [135], [168]	2005	Video Surveillance	15	[169]
ViSOR [170]	2005	Video Surveillance	-	[171]
IXMAS	2006	Multi-camera	13	[172]

CASIA	2007	Behaviour Analysis	8	[173]
UCF Aerial Action	2007	Aerial viewpoint	9	[174]
UCF-ARG	2008	Multi-view (Aerial-Roof-Ground)	10	[175]
UCF sports action [119], [120]	2008	Sports, Annotated	10	[176]
UIUC action [177]	2008	Sports, Annotated	14	[178]
i3DPost multi-view [179]	2009	3D volumes	13	[180]
URADL [181]	2009	Daily Actions	10	[182]
Collective Activity Dataset [183]	2009	Collective Activity, Real World Data	5	[184]
BEHAVE [114]	2010	Activity Benchmark	10	[185]
MuHAVi [186], [187]	2010	Silhouette based methods, Multiview	17	[188]
UT-interaction [189], [190]	2010	Person-Person Actions	6	[191]
UT-tower [192], [193]	2010	Action Recognition at a Distance	9	[191]
UCR-Videoweb [189]	2010	Non-verbal communication analysis, Person-Person Actions	9	[191]
VIRAT video [194]	2011	Realistic video surveillance	12	[195]
MINTA [196]	2011	Kitchen Activities, Humanoid Robot Viewpoint, Intention-Activity-motion primitives distinction	*9(intention), 6(activity), 60(motion)*	[197]
KIT Robo-Kitchen Activity Dataset [127]	2011	Kitchen Activities, Humanoid Robot Viewpoint, Multi-view	14	[198]

Table 2.4 Evolution of Modern datasets

Dataset	Year	Focus	Method	Annotation	Classes	Link
Hollywood [150]	2008	Daily Actions	Movies, Script Alignment	Frame Range /Sequence level	8	[199]
Hollywood2 [200]	2009	Daily Actions	Movies, Script Alignment	Frame Range/Sequence level	20	[201]
UCF11 [202]	2009	Actions in the Wild	YouTube ,Manual Annotation	Sequence level	11	[203]
High Five [204]	2010	Person-Person Actions	TV Shows, Manual Annotation	Sequence level + upper body, head orientation, interaction label	5	[205]

Olympic Sports [121]	2010	Sports Actions, Non-periodic Actions	YouTube, AMT Annotation	Sequence level	16	[206]
UCF50 [207]	2010	Actions in the Wild	YouTube, Manual Annotation	Sequence level	50	[208]
HMDB [111]	2011	Benchmark Dataset	YouTube and Movies, AMT Annotation	Sequence level + meta tags	51	[209]
UCF101 [109]	2012	Benchmark Dataset	YouTube, Manual Annotation	Sequence level	101	[210]
MPII Cooking [143]	2012	Cooking Actions, Fine Grained Actions, Activities of Daily Living	Semi-Scripted (Recipe as Script)	Frame Range + pose	65	[211]
MPII Cooking Composite [212]	2012	Cooking Actions, Composite Actions	Semi-Scripted (Recipe as Script)	Frame Range + ingredient, tool, container labels, script data	41	[213]
ASLAN [161]	2012	Action Similarity Metric	Search Query	Sequence Level	432 **	[214]
Hollywood2Tubes* [215]	2013	Action Localization, Point Localization	Manual Annotation	Point Annotation, Bounding Boxes	20	[216]
YouCook [217]	2013	Cooking Actions, Summarization	YouTube, AMT Annotation	Frame Level (object, actor)	7	[218]
THUMOS'13* [219]	2013	Benchmark	Manual Annotation	Sequence level + bounding boxes, low level attributes	101	[220]
JHMDB* [221]	2013	Joint Annotated using 2D computer Model	AMT Annotation	Joint annotations	21	[222]
Breakfast Actions [223]	2014	Breakfast Actions, Cooking Actions	Semi Scripted (Recipe as Script), Manual Annotation	Frame Range	10	[224]
THUMOS'14* [219]	2014	Benchmark Dataset	YouTube, Manual Annotation	Sequence level	102	[225]
Sports1M [50]	2014	Sports Actions	YouTube ,YouTube Topics API, Search Query	Sequence level	487	[226]

THUMOS'15 [156]	2015	Benchmark Dataset	YouTube, Manual Annotation	Sequence level	102	[227]
Crêpe Dataset [228]	2015	Cooking Actions	Scripted	Dense Frame Level (action, activity, occlusion) +bounding boxes	9	[229]
SVW [230]	2015	Sports Videos, Smartphone Camera View	Crowdsourced	Sequence level/ Frame Range + bounding boxes	44	[231]
ActivityNet [108]	2015	Trimmed and Untrimmed actions, Web	Search Query collection, Manual Verification and Annotation	Sequence level/ Frame Range	203	[232]
MPII Cooking 2 [233]	2015	Cooking Actions, Fine Grained Actions, Composite Actions	Semi-Scripted, AMT	Frame Range + pose, hand annotations, ingredient, tool, container labels, script data	67	[234]
MERL Shopping Dataset [53]	2016	Shopping activities	Surveillance Videos	Frame Range	5	[235]
Charades [146]	2016	Everyday Actions, Home Videos, Person-object	Hollywood in Homes Crowdsourcing	Frame Range + Object labels	157	[236]
MultiTHUMOS [237]	2017	Action Localization,	YouTube, Datang2 Annotation	Dense Frame level	102	[238]
Something Something [147]	2017	Fine Grained Labelling, Caption Template, Levels of Label Granularity	Hollywood in Homes Crowdsourcing	Sequence Level + Object Labels	174	[239]
DALY [145]	2017	Daily Actions, Spatiotemporal Localization,	Search Query, Manual Annotation	Frame Range + Bounding Boxes (actor head, object), Pose (upper body)	10	[240]

AVA [103]	2017	Exhaustive Annotation, Atomic Actions	Movies, Hybrid, Faster RCNN +Manual Annotation	Dense Frame Level + Bounding Boxes (object)	80	[241]
A2D [242]	2017	Actor-Action Correspondence	Search Query	Dense Frame Level + Pixel Level Spatial Masks	9	[243]
Kinetics400 [110]	2017	Human Focused, Benchmark,	YouTube, AMT Annotation	Sequence Level	400	[244]
Kinetics600 [104]	2017	Human Focused, Benchmark,	YouTube, AMT Annotation	Sequence Level	600	[244]
Vlog [107]	2017	Lifestyle Vlogs, Implicit Tagging, Daily Actions, Person-Object	YouTube, Search Query, Manual Annotation	Sequence Level + Attribute Tags, Bounding Boxes	30	[245]
YouCook2 [246]	2018	Cooking Actions, Instructional Videos, Procedure Segmentation Task	YouTube, Manual Annotation	Frame Range (Sentences as Labels)	-	[247]
SoccerNet [123]	2018	Soccer Actions	Sports Broadcasts, Annotation by Commentary Mining	Action Points	3	[248]
MLB YouTube [249]	2018	Fine Grained Activity, Baseball Videos, Overlapping Multilabel Annotations	YouTube, Manual Annotation	Sequence Level/ Dense Frame	9	[250]
STAIR Actions [105]	2018	Fine Grained Activity, Paired Actions	YouTube Home Videos, Search Query, Manual Annotation	Sequence Level	100	[251]
Baseball (BBDB) [252]	2018	Fine Grained Activity, Baseball Videos	Sports Broadcasts, Annotation by Commentary Mining	Frame Range	30	[253]
Moments in Time [115]	2018	Event Detection , Moments, Benchmark	Web Videos, Search Query, AMT Annotation	Sequence Level (short sequence)	339	[254]

HACS [255]	2019	Temporal Localization	YouTube, Search Query, Manual Annotation	Sequence level/ Frame Range	203	[256]
COIN [257]	2019	Instructional Video, Hierarchy of Actions (Domain-Task-Step)	YouTube, Manual Annotation	Frame Range	180	[258]
Mining YouTube [117]	2019	Automatic Extraction of Classes, Cooking Actions	YouTube, Search Query, Text Mining, YouTube Captions, Manual and Automatic Annotation	Frame Range	900	-

* Supplemental Dataset: Provides only additional features for pre-existing dataset.

** Complex Classes (very few examples per class)

Table 2.5 Evolution of Egocentric datasets

Dataset	Year	Focus	Method	Annotation	Classes	Link
GTEA [152]	2011	Head Mounted Camera	Manual Annotation of Unscripted Video	Frame Range	7	[259]
ADL [144]	2012	Chest Mounted Camera, Activities of Daily Living, Person-Object Actions	Semi-scripted	Frame Range + Bounding Boxes (object)	18	[144]
GTEA Gaze [260]	2012	Head Mounted Camera	Manual Annotation of Unscripted Video	Frame Range +Gaze Data	25	[259]
BEOID [261]	2014	Head Mounted, Person- Object Interactions	Scripted	Sequence Level / Frame Range (actions, objects)	34	[262]
DogCentric Activity [263]	2014	Dog Mounted Camera, Animal-Human Actions	Manual Annotation of Unscripted Video	Sequence Level	10	[264]
GTEA Gaze+ [265]	2015	Head Mounted Camera, Cooking Actions	Semi Scripted (Recipe as Script)	Frame Range + Gaze Data, scripts	44	[259]

MILADL [266]	2015	Paired Wrist and Head Mounted Camera, ADL	Semi-Scripted, Manual Annotation	Frame Range	23	[267]
IIIT Extreme Sports [268]	2017	Head Extreme Sports Actions	YouTube	Frame Range	18	[269]
Charades Ego [270]	2018	Paired Egocentric and Third Person Actions	Hollywood in Homes, Crowdsourcing	Frame Range + Object labels	157	[236]
Epic Kitchens [116]	2018	Head Mounted Camera, Kitchen Actions	Crowdsourcing, AMT Annotation	Frame Range + Bounding Boxes (object)	125	[271]
EGTEA [130]	2018	Head Mounted Camera, Cooking Actions	Semi Scripted (Recipe as Script)	Frame Range + Gaze Data, hand mask, scripts	106	[259]
FPVO [272]	2019	Chest Mounted Camera, Office Actions	Manual Annotation of Unscripted Video	Frame Range	20	[273]

2.4.4. Image Datasets

Even though the temporal context plays a pivotal role in the identification of images, it is still possible to recognize certain actions only from static images [274]. To this end, several datasets consisting of labelled still images have been constructed [275] though this has received less interest than video datasets. Some of the image datasets are listed in Table 2.6.

Table 2.6 Evolution of Image datasets.

Dataset	Year	Focus	Method	Classes	Link
Willow [274]	2010	Still Images	Flickr Search Query	7	[276]
PPMI [277]	2010	Person-Object Interactions, Musical Instruments	Manual	12	[278]
Stanford40 [279]	2011	Still Images	Google, Bing, Flickr Search Query	40	[280]
TUHOI [113]	2014	Person Object Interaction	Crowdsourcing, Crowdflower	2974	-
HICO [112]	2015	Person Object Interaction	Flickr Search Query	600	[281]
BU- Action [282]	2017	UCF101, ActivityNet Classes	Google, Flickr Search Query	101 (BU101-F), 101 (BU101-UF), 203 (BU203)	[283]
HICO-DET* [284]	2018	Person-Object Linkage, Dense Annotation with Bounding Boxes	AMT	-	[281]

* Supplemental Dataset: Provides only additional features for pre-existing dataset.

2.5. Dataset used for Training and Evaluation

Throughout our experimental setup we have used several datasets with different classed and modality. A comprehensive list of the datasets used in our experiments is shown in Table 2.7.

Table 2.7 Dataset used for Training and Evaluation in this thesis

Sr. No	Dataset	Description	Approach where dataset used
1	Own Activity Dataset	This dataset contains the static human activities of the video i.e. seating and 6 dynamic activities namely walking, running, jogging, boxing, slap, jogging in different directions. These videos are taken in a substantial indoor environment.	Multi-view human activity recognition framework using multiple features
2	KTH Actions Dataset [285]	This data set contains 6 types of human behaviour's (walking, jogging, running, boxing, hand waving, one-hand pat) that have been performed several times by 25 people in four different scenes. The database contains 2391 sequences.	Multi-view human activity recognition framework using multiple features

		spatial resolution of 160*120 pixels and has an average of 4 seconds	
3	i3DPost Dataset [286]	8 people performing 13 actions (walking, running, jumping, bending, hand-waving, jumping in place, sitting-stand up, running-falling, walking-sitting, running-jumping-walking, handshaking, pulling, and facial-expressions) The actors have different body sizes, clothing and are of different sex, nationality, etc.	Multi-view human activity recognition framework using multiple features
4	MSR action recognition [287]	Contains 5 types of human actions (Standing, Hand-waving, Jumping, Hand-clapping, Boxing) Each video is of low resolution 320 x 240 and frame rate 15 frames per second.	Multi-view human activity recognition framework using multiple features
5	WVU multi-view [288]	Dataset includes different activities hand waving, clapping, jumping, jogging, bowling, throwing, pickup, and kicking. For each view, action sequences performed by different subjects are provided.	Multi-view human activity recognition framework using multiple features
6	NTU-RGB+D Dataset [289]	contains 57K videos for 60 activities classes performed by 40 distinct subjects and 80 viewpoints. The resolution of each depth frame is 512 × 424 pixels For this evaluation, the training and testing sets have 40,320 and 16,560 samples, respectively.	Depth based enlarged temporal dimension of 3D deep convolutional network for activity recognition and Dual Stream HAR model exploiting Residual-CNN
7	MSRAction3D Dataset [290]	consists of 20 activity types performed by 10 subjects. Every activity is performed by every subject 2 or 3 times. There are 567 depth sequences of resolution 640 x 240 pixels in total	Depth based enlarged temporal dimension of 3D deep convolutional network for activity recognition
8	MSRDailyActivity3D Dataset [126]	contain 16 activities: drink, eat, walking, read book, write on paper, use laptop, cheer up, use vacuum cleaner, sit still, toss paper, play guitar, play game, lay down on sofa, sit down, stand up, and call cell phone. Ten different subjects perform each activity two times, once in standing and the other in sitting position. There are a total of 320 depth sequences in the dataset.	Depth based enlarged temporal dimension of 3D deep convolutional network for activity recognition And Combing CNN Streams of Dynamic Image and Depth Data for Action Recognition

9	UCF101 [210]	Consists of total 13320 video Different action videos are grouped into 25 groups Each group contains 4-7 videos of an action. Five types of action categories: 1- Body-Motion Only 2- Human- Object Interaction 3- Human-Human Interaction 4- Playing Musical Instruments 5- Sports	Dual Stream HAR model exploiting Residual-CNN and Human Activity Recognition using CRNN
10	HMBD-51 [111]	Consists of around 7,000 manually annotated clips. The dataset has 51 action categories. The actions can be grouped into five types: 1) General facial actions 2) Facial actions with object manipulation 3) General body movements 4) Body movements with object interaction 5) Body movements for human interaction.	Dual Stream HAR model exploiting Residual-CNN
11	UTD MHAD dataset [291]	27 different actions such as basketball shoot, squats, boxing, bowling, drawing triangle etc. are captured in indoor environment by 8 different subjects consisting of 4 male and 4 females. All the different actions are performed twice by each subject resulting in total of 861 sequences. The data captured in UTD MHAD dataset is of four different types RGB, Depth, Skeletal and Inertial data. We have used RGB and depth data only.	Combing CNN Streams of Dynamic Image and Depth Data for Action Recognition
12	CAD-60 dataset [292]	12 Various daily indoor activities such as relaxing on the couch, talking on the couch, rinsing mouth, brushing teeth, wearing lens, talking on phone, working on computer, chopping vegetable, stirring etc. resulting in 60 videos are recorded.	Combing CNN Streams of Dynamic Image and Depth Data for Action Recognition

2.6. Performance Measures

In this section, the brief descriptions of parameters used to evaluate the performance measures of human activity recognition are discussed as follows:

A confusion matrix of binary classification is a two by two table formed by counting of the number of the four outcomes of a binary classifier.

Table 2.8 Four outcomes of a binary classifier

		Predicted Class	
		Positive (1)	Negative (0)
Observed Class	Positive (1)	TP	FN
	Negative (0)	FP	TN

In above table 2.8, TP corresponds to true positive instances, FN corresponds to false negative instances, FP corresponds to false positive instances and TN represents true negative instances.

2.6.1. Accuracy and Error

Accuracy is defined as number of correct predictions made by the model over all the predictions made. It is defined as follows.

$$Accuracy = \frac{\text{TP+TN}}{\text{TP+FP+FN+TN}} \text{ x 100 (in Percentage)} \qquad (2.1)$$

$$\text{Error} = 100 - \text{Accuracy (in Percentage)} \qquad (2.2)$$

2.6.2. Recall, Precision & Specificity

The precision, recall and specificity parameters are calculated as follows

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \qquad (2.3)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \qquad (2.4)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{FP+TN}} \qquad (2.5)$$

Here TPR represents true positives rate, FNR represents false negatives rate, FPR represents false positives rate and TNR represents true negative rate.

2.6.3. F-Score Measure

Let we have n activity classes for classification C_1, C_2 . . . . . C_n

Precision and Recall for individual classes may be calculated using above equations.

Let precision and recall for classification of individual classes be P_1, P_2, ………. P_n and R_1, R_2, …….. R_n.

Macro Precision (P_r) and Recall (R_e) can be calculate as follow

$$P_r = (P_1 + P_2 + \ldots\ldots\ldots\ldots + P_n) / n \qquad (2.6)$$

$$R_e = (R_1 + R_2 + \ldots\ldots\ldots\ldots + R_n) / n \qquad (2.7)$$

Finally, the F-Score for the proposed framework has been calculated as-

$$\text{F-Score} = \frac{2 \text{ x } Pr \text{ x } Re}{Pr + Re} \qquad (2.8)$$

2.7. Conclusion

This chapter presented a theoretical background of the video surveillance system. In this chapter, an overview of human activity recognition techniques using conventional and deep learning approaches has been presented. The complete overview has been covered in two parts – the first part covers different approaches for human activity recognition and the second part cover evolution of modern datasets for human activity recognition.

Firstly, a detailed survey of different HAR method using conventional and deep learning based approaches has been presented. The conventional approaches include space time based, appearance based and other categorizations like LBP, Soft Computing, etc. while Deep learning based approaches includes Vision Based, Depth sensor based categorizations. A short description of some deep learning approaches like Convolutional Neural Network, Recurrent Neural Network etc. has been presented. Further, based on the survey of various human activity approaches, this chapter presents Research Gaps in the area of human activity recognition in video surveillance systems.

Secondly, a comprehensive survey on the evolution of early and modern dataset for human activity recognition has been presented. Further, a detailed list of the various dataset used in this thesis for training and evaluation of the various proposed models has been listed along with different performance measure used in this thesis.