Vision Based Human Activity Recognition: A Comprehensive Review of Methods & Techniques

Palak Girdhar[1], Prashant Johri[2], Deepali Virmani[3]

[1]School of Computer Science & Engineering, Galgotias University, Greater Noida, UP
[2]School of Computer Science & Engineering, Galgotias University, Greater Noida, UP
[1,3]Department of Computer Science & Engineering, Bhagwan Parshuram Institute of Technology, Delhi

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 28 April 2021

Abstract
Human Activity Recognition (HAR) plays an important role in various domains. There are four major phases of HAR: Data Collection, Pre-Processing, Feature Extraction and Training. Out of these, Pre-processing and Feature Extraction requires massive attention, as these are the building block for the training phase. In this paper, an extensive and thorough review of the state-of-the-art methods is given along with identification of challenging areas of HAR ecosystem. Primarily, trends in research outflowing in the area of Vision based approach is effectively studied in the present review. Further, a consideration with state-of-the-art applications and the importance of handcrafted and learning based features is emphasized. The review presented in this paper, gives the details of available training datasets of HAR and also discusses the popular publicly available datasets. The readers will be benefitted with a comprehensive review of the extensive work done in the field of HAR and its Vision based approaches.

Keywords: HAR, Sensor and Vision-based techniques, Local and Global Feature Representation, HAR Datasets

1.0 Introduction

Human Activity Recognition (HAR) is gaining more popularity today due to its numerous applications present in real world like in the area of Human Computer Interaction (HCI), Ambient Assisted Living (AAL), Intelligent Video Surveillance, Human -Robot interaction, Human Behaviour understanding (Vrigkas, Nikou & Kakadiaris, 2015).

Furthermore, the use of technology in activity recognition is incomplete without a detailed analysis of the human activity involved in various captured frames. In order to observe *human activity (collection of actions),* it requires an efficient analysis which can be done at different levels. The division is purely on the basis of complexity and time span of the action like from a simple gesture to a group activity. The human activities are categorized into the following levels

- Gesture - A gesture is a primitive movement of any body part to convey some information. It can be a small hand movement or just the facial expression. The duration for such activities is very small.
- Atomic Action - it is a simple activity (involves several gestures) poses by a human being. Examples are like: jogging, running, swimming.
- Interaction - it is an activity between human and another agent. The other agent involved in interaction can be a human being or any object. If the interaction is in between human and human then it is termed as human-human interaction. Examples of such interactions are: shaking hands, hugging each other. And if one of the agents in interaction is any object then it is human to object interaction like a person is using his laptop.
- Group Actions- it is a complex activity which involves more than two human and objects. Example of group activities are: playing a game in a team like football, playing cards in a group, group fighting, parade etc.

From figure 1 it is clearly understood, as we move around the axis, the complexity level of the human activities also gets increased. Hence, difficulty to automate the process also reaches a certain extent.

7383

Figure 1: Levels of Human Activities

In recent times, sensor-based technology has achieved a great and exceptional performance in various aspects like computational power, size, cost and accuracy. Due to these achievements, low power and small sized sensors are able to integrate with smartphones and other portable devices. Besides, vision- based technology is also gaining more attention. Evolution of Closed-Circuit Television (CCTV) provide better video quality, lower cost setup and enable continuous monitoring. Both sensors based and vision-based technology has its own applications. HAR comprises of 4 major phases (Sargano, Angelov &Habib,2017): 1) Data Collection 2) Pre-Processing 3) Feature Extraction and Training 4) Activity Recognition.

With the rapid growth in the technology, there is a lesser demand of human intervention in the phases of HAR. However, the performance of these systems is highly data dependent. Therefore, HAR frameworks are dependent on the sensing devices for the *data collection*.

Further, after data collection, the next important step is to pre-process the data. Pre-processing is required before feeding the data into any training algorithm. It is considered to be most important part of the process. Performance of any HAR system majorly depends on the pre-processing task. The pre-processing is quite tedious in Vision-based HAR system. Due to change in view point, occluded background, lighting condition makes the HAR system more challenging. Therefore, pre-processing plays a crucial and important role in the performance of HAR systems. Segmentation is the process which focus on the target object from the given set of images. It can be further classified into background removal/subtraction (Babee, Dinh & Rigoll,2018; Bouwmans,2014) and foreground extraction (Allili, Bouguila, Ziou,2007; Bouwmans, 2014). Background subtraction is popular for extracting the moving region from the sequence of frames captured, where the background is not dynamic. It is most suitable for the regions where the camera position and angle are fixed. It also works on the pixel-by-pixel differencing between the current image and the reference background image. However, it is sensitive to changes of lighting condition, clutter background, occlusion etc. Statistical model (Bouwmans, 2014), optical flow model (Vishwakarma, Agrawal,2013) works on the principle of background modelling. Foreground extraction is the process of focusing on moving object present in the image or video. It is very challenging as compared to extraction recorded with static camera, as in this scenario background and foreground keeps on changing. Optical flow (Vishwakarma, Agrawal,2013), temporal differencing (Vishwakarma, Agrawal,2013), Hidden Markov modelling (Hu, Xie, Zeng & Maybank,2011) are the common techniques that are used to calculate the foreground object. The third important and integral part of the HAR system is Feature Extraction. It is discussed in the section 3 in more detail. Taxonomy of Vision-based human activity recognition is shown in Figure 2.

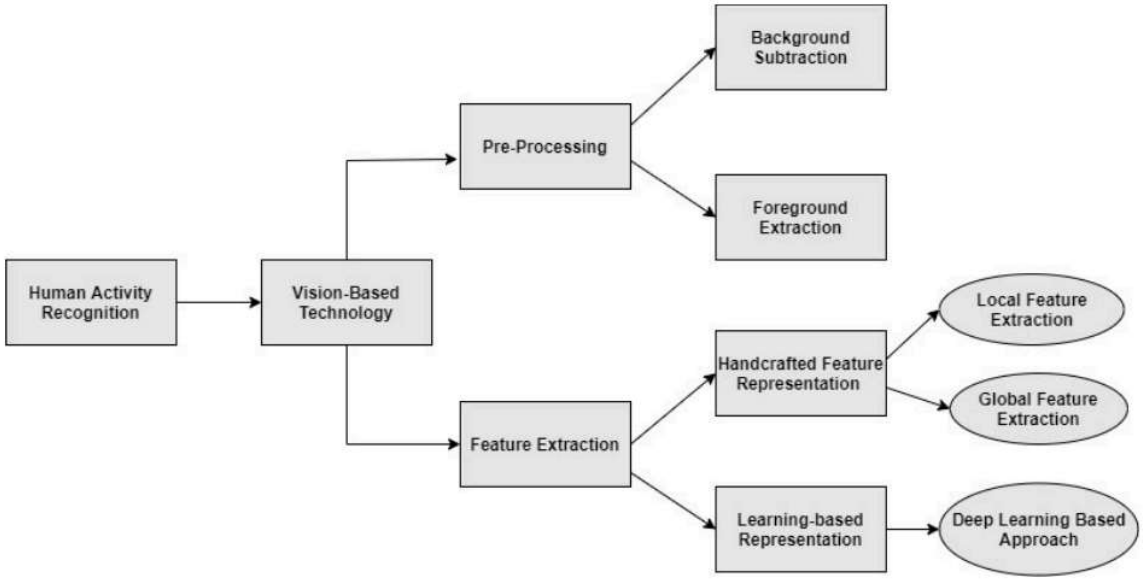

Figure 2: Vision-Based Human Activity Recognition

Outline: The paper extends with related work in section-2. Section 3 focuses on handcrafted feature extraction methods and learning-based features techniques. Furthermore, Section-4 gives the summary of available datasets for the human action recognition.

2.0 Related Work

There are numerous surveys present on HAR in the literature (Vrigkas, Nikou & Kakadiaris, 2015; Sargano, Angelov & Habib, 2017; Vishwakarma & Agrawal, 2013; Popoola & Wang, 2012; Borges, Conci & Cavallaro, 2013). These surveys provide a major insight to recent trends in the field of automated human activity recognition. The listed surveys emphasize on different aspects of any HAR system. Few of them focuses on spatial temporal information, few examined HAR approaches according to the complexity level of features extracted. Some of the popular surveys are listed in the Table 1. It gives insight to the popular surveys from 2012-2020.

Table 1: Summary of Popular Surveys

Year	Focused On	Reference
2012	Abnormal human behavior recognition	(Popoola & Wang, 2012)
2013	Video -based human behavior understanding	(Borges, Conci & Cavallaro, 2013)
2013	Human activity recognition and understanding	(Vishwakarma & Agrawal, 2013)
2014	Intelligent video surveillance systems for public spaces	(Zablocki, Gosciewska, Frejlichowski & Hofman, 2014)
2014	Human action recognition using vision-based techniques	(Hassan, Ahmad, Liaqat, Farooq, Ali & Hassan, 2014
2014	Human activity recognition from 3D data	(Aggarwal & Xia, 2014)
2015	Unimodal and multimodal approaches for human behavior understanding	(Vrigkas, Nikou & Kakadiaris,

		2015)
2016	Spatio-temporal based model for human action recognition from skeleton data	(Song, Lan,Xing, Zeng & Liu, 2016)
2017	Video data analysis using traditional and deep learning-based approaches	(Herath, Harandi & Porikli, 2017)
2018	Action detection, recognition and evaluation through captured motion data	(Patrona, Chatizitofis, Zarpalas & Daras, 2018)
2019	Sensor based human activity recognition	(Zhang, Zhang, Zhong, Lei, Yang, Du & Chen, 2019)
2020	Vision based human action recognition	(Jegham, Khalifa, Alouani & Mahjoub, 2020)

2.1. State-of-the-Art Review

For the development of any HAR system, data collection is the most basic and most important primitive to work upon. *For the data collection many devices are available.* Data can be collected via sensor-based or via vision-based devices. Sensor based devices are small enough that can be integrated with any of the devices that are used on daily basis like smartphones, watches etc. On the other hand, in applications like automatic surveillance system vision-based devices are more useful. Devices like CCTV are efficient to capture/monitor the unforeseen activities from a distance. So, this subsection gives an overview of important characteristics of sensor and vision-based data collection. Table 2 gives the insight of the popular surveys present in the field of Vision based approach.

- *Sensor based HAR System*

Sensor based HAR is very popular in real world due to its applications like in healthcare, smart homes and many more. In sensor based HAR, a sensor device is attached to human body to collect the activities of a person involved, like a simple Fitbit device monitors the steps you are walking. It collects the information and, with that collected information, it generates the results like a person's pulse rate and other vital features of health monitoring. Sometimes sensors are attached in the surroundings to collect the information. For example, in smart homes, a temperature sensing device can help in early fire detection and raise the alarm for taking the safety measures at appropriate time.

- *Vision based HAR System*

Vision based system relies on visual sensing of the environment like CCTV to record and monitor the human activities. This approach works well if the quality of the captured image is good. Camera quality, image resolution, lighting conditions, illumination change are other responsible parameters for the image quality. The gathered information is a collection of still images or some un processed video files. To automate the human activity recognition, it is required to pre-process this collected data. Numerous pre-processing techniques are available in the literature (Vishwakarma & Agrawal, 2013; Aggarwal & Xia,2014). Some initial steps include human detection, modelling and segmentation on the basis of activity, activity classification and tracking (Vishwakarma & Agrawal, 2013).

From last few years, vision based HAR system has gained much attention in the real-world applications. To ensure the safety and monitor the activities, CCTVs are installed everywhere. Applications like child day care to airport surveillance, such systems are in a huge demand. Table 2 represents some of the literature work present in the field of HAR, based on vision sensing technology.

The accuracy of the system depends on the input that is supplied to the system. Input to any vision based HAR system, are the collected frames. These collected frames are the sequence of RGB images. Some traditional system works on RGB data (Sargano, Angelov & Habib, 2017; Wu, Ma, Zang, Wang & Li, 2017) and some on RGB-D data (Wu, Ma, Zang, Wang & Li, 2017). The system with RGB data as input are less expensive and less accurate too whereas RGB-D data provides results with good accuracy. RGB-D is the combination of RGB and its corresponding depth image. Each pixel represents the distance between the image plane and object in the corresponding RGB image. The reason for achieving good accuracy with RGB-D data is that it is robust against lighting conditions, illumination effect, colour of the image, texture and even works well in the dark environment.

Some of the available RGB-D datasets in this field are: HuDA Act dataset (Ni, Wang & Moulin,2011), MV-TJU (Zhang, Lin, Nie, Chaison, Wong & Kankanhalli, 2015), SDUFall (Apicella, Snidaro, 2021), CAD-60 (Faria, Premebida & Nunes, 2014), MSR Daily Activity 3D (Ali, Moftah & Youssif, 2018) data set gives more insight to research direction.

Table 2: Vision Based Surveys

Year	Main contribution	Reference
2016	• Focused on knowledge based HAR • Mainly focused on HAR methods for video streams.	(Onofri, Soda, Pechenizkiy & Iannello, 2016)
2017	• Focused on a comprehensive review for recognizing HAR. • Shows various handcrafted and deep learning-based approaches for feature extraction. .	(Herath, Harandi & Porikli, 2017)
2018	• Focused on integration of explicit knowledge and vision-based data for better outcome.	(Souza Alves, Oliveira, Sanin & Szczerbicki, 2018)
2019	• Focused on depth video processing to track depth silhouettes • Use skeleton joints to monitor daily activities of people.	(Kim, Jalal & Mahmood, 2019)
2020	• Presented a comprehensive survey for HAR and methods used for the feature extraction purpose. • Focused on the data selection for the better accuracy attainment.	(Jegham, Khalifa, Alouani & Mahjoub, 2020)

3.0Human Action Feature Representation Methods

Feature extraction is an important phase for the success of human action recognition system. The aim of this step is to extract some useful information from the image so that a good descriptor can be found. The collected data (either through sensors or CCTVs) undergoes with the pre-processing task first. Once the data is pre-processed, it is further required to extract important features from the processed data for effective classification. Different approaches have been used in the literature for

the feature extraction purpose. In this section, Handcrafted features (based on local and Global feature extraction) and learning based features is studied. This section also presents an insight to importance of feature detector and descriptors and also discuss the widely used methods for feature descriptors used in global representations and local representation approaches.

3.1 Handcrafted Features

Feature Extraction is a process of extracting some useful information from the given set of large raw data. It is the process of finding some useful information by reducing the dimensions of the data. Various approaches have been used in the literature. Handcrafted feature-based representation is a one of the machine learning based traditional approach. This method has widely used in the literature for long time. It gave a remarkable result in the area of human activity recognition applications (Dang, Min, Wang, Piran, Lee & Moon, 2020).

However, this approach is time consuming, due to manual intervention. Figure 3 gives the clarity of the process. For the correct recognition, set of features are selected and validated manually by the experts. Before the emergence of deep learning methods, most of the action recognition frameworks followed this approach which first extracts the features in spatial and motion domain and then fed this information into some specific classifier like Support Vector Machine (SVM) (Jegham, Khalifa, Alouani & Mahjoub, 2020; Dang, Min, Wang, Piran, Lee & Moon, 2020) for the classification purpose. Handcrafted feature approach is further categories into two classes. Global Feature Extraction approach (Dang, Min, Wang, Piran, Lee & Moon, 2020) considers the image as a one and doesn't focus on the individual parts of the image whereas, local feature extraction approach (Dang, Min, Wang, Piran, Lee & Moon, 2020) focuses on the local areas of the image for the better results.

- *Global Feature Extraction*

Global features consider the given image a s a single vector. It describes the image as a whole rather than focusing on the individual part of the image. It includes contour representations, shape descriptors and texture features. Shape metrices, invariant moments, Histogram oriented Gradients (HOG), Bag of Words, Motion Boundary Histogram (MBH) (Sargano, Angelov & Habib, 2017) are some common global feature descriptors present. Global features have few limitations in conditions like illumination variation, sensitivity to noise etc. In such conditions it is difficult to extract the relevant information. Therefore, it is not suitable for some of the applications. Hence, Local feature extraction methods which focuses on the Interest Points (IPs) are adopted.

- *Local Feature Extraction*

Instead of taking the whole image, this method focuses on the Local feature of the image. These methods use local descriptors to represent an input image. It concentrates on local patches of interest. Finding the point of interest is the foremost step in any recognition task. It gives more clarity of those areas which are most important for describing the object. Feature detectors helps in finding that area. And Feature descriptors helps to encode the information of the neighbourhood of the selected point. It focuses only on those regions which carry some information. Most of the real-world applications deal with scale invariance, occlusion and deformation. It is required that feature detectors and descriptors can handle the mentioned challenges. Scale -invariant feature transform (SIFT), speed-up robust feature (SURF) are the most popular local feature descriptors used in the literature (Sargano, Anglelov & Habib, 2017; Herath, Harandi & Porikli, 2017).

3.2 Learning-based Features

Handcrafted features focus on the pixel level of the image or of the video. It mainly focuses on the spatial shape or the temporal motion. The more advanced approaches have the capabilities to learn automatically from the raw data. Figure 4 shows that it supports end-to-end learning. The performance of any HAR system depends on the efficient representation of extracted features. Unlike, in

handcrafted feature representation methods, feature detectors and feature descriptors are used whereas in learning-based representation, the system has the capability to extract and learn from the raw data.

There are two approaches present in this category: Non-Deep learning-based approaches and Deep-learning based approaches (Herath, Harandi & Porikli, 2017; Zhu, Shao, Xie, Fang, 2016). Genetic Programming and Dictionary based learning falls under the first category, which is not discussed here. Whereas generative and supervised learning-based approaches falls under Deep Learning based category, discussed further.

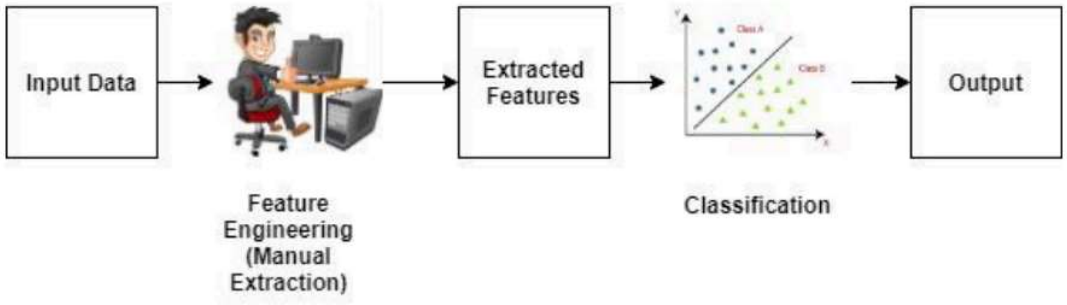

Figure 3: Handcrafted Feature Representation

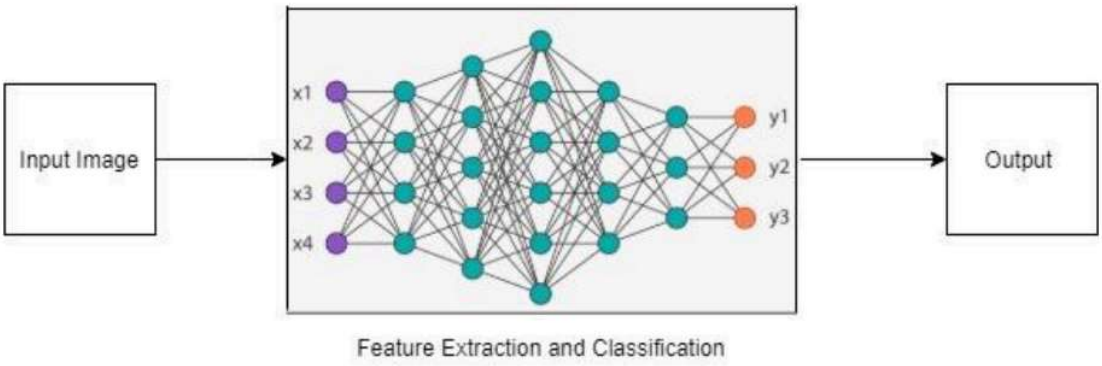

Figure 4: Learning-based Feature Representation

3.2.1 Deep Learning-based approaches

Over the last few years, Deep Learning has shown a remarkable result in the field of computer vision. It is an emerging area of machine learning which provides a deep level of representation and abstraction. Deep learning-based techniques have ability to work upon the raw data and it has the capability to extract and represent features automatically. These techniques use trainable feature extractor and multiple hidden layers for deep processing of the data. Figure 4 shows how it supports end to end learning. Deep learning-based approaches are further divided into Generative and Supervised Models (Herath, Harandi & Porikli, 2017; Zhu, Shao, Xie, Fang, 2016) which are discussed in detail.

- *Generative Models*

Generative models work on the principle of reconstruction. It reconstructs the dataset on the basis of some probabilistic model (Dang, Min, Wang, Piran, Lee & Moon, 2020). Unlike, Supervised learning, it does not require any target class during the training phase. This model is useful when there is a scarcity of labelled training data. Due to this limitation, Deep Belief Networks (DBN) (; Zhu, Shao,

Xie, Fang, 2016) came into existence. This model was relatively faster than traditional generative models. Training of these networks performed by Restricted Boltzmann Machine (RBM) (Dang, Min, Wang, Piran, Lee & Moon, 2020; Zhu, Shao, Xie, Fang, 2016) in a layer-by-layer fashion. It uses the concept of backpropagation. This technique is widely used in the applications like in generation of cartoon characters, in generation of realistic human faces, human posing, face aging and many more.

- *Supervised Models*

Supervised models are trained with the labelled data. During training phase, each training input is labelled with the desired output value. Training set is the pair of input and corresponding output value. Convolutional Neural Network (CNN) has shown the excellence in the field of pattern recognition, image classification and human activity recognition (Ullah, Ahmad, Muhammad, Sajjad & Basik, 2017). CNN takes the input and pass the input to the densely connected layers and with each iteration the weights of the hidden layers are adjusted to give the proper classification results. The structure of CNN has convolutional layer, pooling layer and fully connected layer. Some hybrid models are also present. These kinds of models use the characteristics of both generative as well as supervised models. For example, to achieve the proper classification results, reconstruction rule can be applied for.

4.0Popular Dataset Available

For the development of any HAR system, dataset selection and collection are a very important aspect. A dataset with a lesser number of angle (viewpoints), subjects involved, occluded scenario, lighting condition results in a limited ability for the recognition. It is analyzed that no single dataset is suitable for all types of activity monitoring because it highly depends on the application and type of the data. Table 2 gives the summary of popular datasets that are publicly available. It gives more clarity of the type of dataset and also number of classes for the recognition purpose has been listed.

Table 2: Popular Human Action Recognition Datasets Available

Dataset	Source	Duration/No. of video files	Type of Dataset	Focused On
KTH (Schuldt, Laptev & Caputo,2004)	Recorded Videos (indoor and outdoor)	600	Human Activity	Six actions (walking, jogging, running, boxing, hand waving and hand clapping) with homogenous background, Static camera
Weizmann (Zelink-Manor & Irani, 2001)	Recorded Videos (indoor and outdoor)	90	Human Activity	bend, jumping jack, jump forward, jump in place, run, gallop sideways, skip, walk, wave one hand and wave both hands.
CAVIAR (Chaquet, Carmona & Fernandez-Caballero,2013)	Recorded Videos (indoor and outdoor)	28	Surveillance	people walking alone, meeting with others, window shopping, entering and exiting shops, fighting and passing out and last, leaving a package in a public place.
ViSOR (Chaquet, Carmona & Fernandez-Caballero,2013**)**	Indoors and outdoors	474	Surveillance (Person Reidentification and Tracking)	Repository for large set of multimedia data and the corresponding annotations.
UT Tower (Ryoo,	Outdoors	109	Aerial View Human	Shake, hug, kick, punch, Push

Chen,Aggarwal & Roy-Chowdhury,2010)			Activity (Low Resolution)	
UCF Aerial Action Dataset (Anuradha & Sairan, 2016)	Outdoors		Surveillance	Walking, Running, Digging, picking up object, kicking, opening car door, closing car door, opening car trunk, closing car trunk.
UCF-Crime Dataset (Girdhar, Johri, Virmani, 2020)	Indoors and outdoors	128 hours of videos	Surveillance	Abuse, Arrest, Arson, Assault, Road Accident, Burglary, Explosion, Fighting, Robbery, Shooting, Stealing, Shoplifting, and Vandalism.
UCF Sports Dataset (Soomro & Zamir, 2014)	Indoors and outdoors	200	Sports Dataset	Diving, golf swinging, kicking Kicking, lifting, horseback riding, running, skating, swinging, walking
i-LIDS (Bloom, Makris & Argyriou, 2012)	Indoors and outdoors	600 Image sequences	Surveillance	Abandoned baggage detection, Parked vehicle detection, Doorway surveillance, Sterile zone monitoring, object tracking
PETS 2010 (Ellis & Ferryman, 2010)	Indoors and outdoors	Nearly 1.6 GB for people tracking and around 1.2 GB for event recognition	Crowd Surveillance	Walking, running, evacuation (rapid dispersion), local dispersion, crowd formation and splitting
VIRAT (Oh, Hoogs, Perera, Cuntoor, Chen, Lee & Desai, 2011)	Outdoors	17	Video Surveillance and supports content- based searching	Person getting into vehicle, getting out of vehicle, Person loading and unloading vehicle, Person opening and closing trunk, Person walking, Person carrying object, Vehicles Parking, Vehicle picking person, Vehicle dropping off person, Person exchanging object from one vehicle to another, Person handing an object to another person, Group of people gathering, Group of people dispersing, Group of people moving
Behave (Blunsden & Fisher, 2010)	Outdoor	9	Human Behavior Recognition	People standing in group, approaching each other, walk together, meet, split, ignore, chase, fight, run together, following
HUMAN4D (Chatzitofis, Saroglou, Boutis, Drakoulis,Zioulis, Subramanyam & Daras, 2020)	Indoors and outdoors	More than 50K Samples	Human centric-multimodal dataset for motion and audio capturing	*Single Person*- Running, jumping jack, bending, punching n kicking, basket-ball dribbling, laying and sitting down, sitting on chair, talking, object dropping and picking, stretching and talking, walking and talking, watching scary movie, in flight safety announcement *Multi-person*- watching football together,

dancing, card trick, whispering, p

5.0Conclusion and Future Work

The popularity of HAR systems makes it necessary to understand its integral components. In this paper, the application areas of HAR are discussed in detail along with the level of human activities in which it can be categorized. This paper has given a comprehensive overview of recent state-of-the-art work in this field. Different approaches have been discussed for the data collection. It also gives insight on the latest surveys present in the field of vision-based sensing technology.

In this paper, it is observed and analysed that the performance of HAR system also depends on the quality of data that is being fed to the model. In existing state of the art surveys and reviews, the importance of RGB and RGB-D datasets is not discussed.

In this paper, the effective use of RGB and RGB-D datasets is discussed in detail. It also presents the importance of feature extraction process, which is an integral part of the four phases of HAR (as discussed in detail in section I). Traditional handcrafted and learning-based approaches have also been discussed. It is observed that there is no single data available that is suitable to all kind of applications. So, this paper provides a careful insight into the well-known publicly available datasets and their detailed description for activity recognition. In future, the present comprehensive view of the tools and techniques of Vision based activity recognition can be investigated more deeply with human activity recognition. This can also be implemented with other soft computing techniques like fuzzy logic, genetic algorithm.

References

1. Allili, M. S., Bouguila, N., & Ziou, D. (2007, May). A robust video foreground segmentation by using generalized gaussian mixture modeling. *In Fourth Canadian conference on computer and robot vision (CRV'07)* (pp. 503-509). IEEE.
2. Aggarwal, J. K., & Xia, L. (2014). Human activity recognition from 3d data: A review. *Pattern Recognition Letters*, 48, 70-80.
3. Apicella, A., & Snidaro, L. Deep Neural Networks for real-time remote fall detection. *Pattern Recognition*
4. Ali, H. H., Moftah, H. M., & Youssif, A. A. (2018). Depth-based human activity recognition: A comparative perspective study on feature extraction. *Future Computing and Informatics Journal*, 3(1), 51-67.
5. Anuradha, K., & Sairam, N. (2016). Spatio-temporal based approaches for human action recognition in static and dynamic background: a survey. *Indian Journal of Science and Technology*, 9(5), 1-12.
6. Bouwmans, T. (2014). Traditional and recent approaches in background modeling for foreground detection: An overview. *Computer science review*, 11, 31-66.
7. Babaee, M., Dinh, D. T., & Rigoll, G. (2018). A deep convolutional neural network for video sequence background subtraction. *Pattern Recognition*, 76, 635-649.
8. Blunsden, S., & Fisher, R. B. (2010). The BEHAVE video dataset: ground truthed video for multi-person behavior classification. *Annals of the BMVA*, 4(1-12), 4.
9. Borges, P. V. K., Conci, N., & Cavallaro, A. (2013). Video-based human behavior understanding: A survey. *IEEE transactions on circuits and systems for video technology,* 23(11), 1993-2008.
10. Bloom, V., Makris, D., & Argyriou, V. (2012, June). G3D: A gaming action dataset and real time action recognition evaluation framework. *In 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (pp. 7-12). IEEE.
11. Chatzitofis, A., Saroglou, L., Boutis, P., Drakoulis, P., Zioulis, N., Subramanyam, S., ... & Daras, P. (2020). HUMAN4D: A Human-Centric Multimodal Dataset for Motions and Immersive Media. *IEEE Access*, 8, 176241-176262.

12. Chaquet, J. M., Carmona, E. J., & Fernández-Caballero, A. (2013). A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, 117(6), 633-659.
13. Dang, L. M., Min, K., Wang, H., Piran, M. J., Lee, C. H., & Moon, H. (2020). Sensor-based and vision-based human activity recognition: A comprehensive survey. *Pattern Recognition*, 108, 107561.
14. de Souza Alves, T., de Oliveira, C. S., Sanin, C., & Szczerbicki, E. (2018). From knowledge-based vision systems to cognitive vision systems: a review. *Procedia Computer Science*, 126, 1855-1864.
15. Ellis, A., & Ferryman, J. (2010, August). PETS2010 and PETS2009 evaluation of results using individual ground truthed single views. In 2010 7th *IEEE international conference on advanced video and signal-based surveillance* (pp. 135-142). IEEE.
16. Faria, D. R., Premebida, C., & Nunes, U. (2014, August). A probabilistic approach for human everyday activities recognition using body motion from RGB-D images. *In The 23rd IEEE international symposium on robot and human interactive communication* (pp. 732-737). IEEE.
17. Girdhar, P., Johri, P., & Virmani, D. (2020). Incept_LSTM: Accession for human activity concession in automatic surveillance. *Journal of Discrete Mathematical Sciences and Cryptography*, 1-15.
18. Hu, W., Xie, N., Li, L., Zeng, X., & Maybank, S. (2011). A survey on visual content-based video indexing and retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews),* 41(6), 797-819.
19. Hassan, M., Ahmad, T., Liaqat, N., Farooq, A., Ali, S. A., & Hassan, S. R. (2014). A review on human actions recognition using vision-based techniques. *Journal of Image and Graphics*, 2(1), 28-32.
20. Herath, S., Harandi, M., & Porikli, F. (2017). Going deeper into action recognition: A survey. *Image and vision computing*, 60, 4-21.
21. Jegham, I., Khalifa, A. B., Alouani, I., & Mahjoub, M. A. (2020). Vision-based human action recognition: An overview and real-world challenges. *Forensic Science International: Digital Investigation*, 32, 200901.
22. Kim, K., Jalal, A., & Mahmood, M. (2019). Vision-based Human Activity recognition system using depth silhouettes: A Smart home system for monitoring the residents. *Journal of Electrical Engineering & Technology*, 14(6), 2567-2573.
23. Ni, B., Wang, G., & Moulin, P. (2011, November). Rgbd-hudaact: A color-depth video database for human daily activity recognition. *In 2011 IEEE international conference on computer vision workshops (ICCV workshops)* (pp. 1147-1153). IEEE.
24. Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C. C., Lee, J. T., ... & Desai, M. (2011, June). A large-scale benchmark dataset for event recognition in surveillance video. In CVPR 2011 (pp. 3153-3160). *IEEE*.
25. Onofri, L., Soda, P., Pechenizkiy, M., & Iannello, G. (2016). A survey on using domain and contextual knowledge for human activity recognition in video streams. *Expert Systems with Applications*, 63, 97-111.
26. Popoola, O. P., & Wang, K. (2012). Video-based abnormal human behavior recognition—A review. *IEEE Transactions on Systems*, Man, and Cybernetics, Part C (Applications and Reviews), 42(6), 865-878.
27. Patrona, F., Chatzitofis, A., Zarpalas, D., & Daras, P. (2018). Motion analysis: Action detection, recognition and evaluation based on motion capture data. *Pattern Recognition*, 76, 612-622.
28. Ryoo, M. S., Chen, C. C., Aggarwal, J. K., & Roy-Chowdhury, A. (2010, August). An overview of contest on semantic description of human activities (sdha) 2010. In International Conference on Pattern Recognition (pp. 270-285). *Springer, Berlin, Heidelberg*.
29. Schuldt, C., Laptev, I., & Caputo, B. (2004, August). Recognizing human actions: a local SVM approach. In Proceedings of the 17th International Conference on *Pattern Recognition, 2004. ICPR 2004*. (Vol. 3, pp. 32-36). IEEE.
30. Soomro, K., & Zamir, A. R. (2014). Action recognition in realistic sports videos. In *Computer vision in sports* (pp. 181-208). Springer, Cham.

31. Song, S., Lan, C., Xing, J., Zeng, W., & Liu, J. (2016). An end-to-end spatio-temporal attention model for human action recognition from skeleton data. *arXiv* preprint arXiv:1611.06067.
32. Sargano, A. B., Angelov, P., & Habib, Z. (2017). A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition. *applied sciences*, 7(1), 110.
33. Ullah, A., Ahmad, J., Muhammad, K., Sajjad, M., & Baik, S. W. (2017). Action recognition in video sequences using deep bi-directional LSTM with CNN features. *IEEE access*, 6, 1155-1166.
34. Vishwakarma, S., & Agrawal, A. (2013). A survey on activity recognition and behavior understanding in video surveillance. *The Visual Computer*, 29(10), 983-1009
35. Vrigkas, M., Nikou, C., & Kakadiaris, I. A. (2015). A review of human activity recognition methods. *Frontiers in Robotics* and AI, 2, 28.
36. Wu, H., Ma, X., Zhang, Z., Wang, H., & Li, Y. (2017). Collecting public RGB-D datasets for human daily activity recognition. *International Journal of Advanced Robotic Systems*, 14(4), 1729881417709079.
37. Zabłocki, M., Gościewska, K., Frejlichowski, D., & Hofman, R. (2014). Intelligent video surveillance systems for public spaces–a survey. *Journal of Theoretical and Applied Computer Science*, 8(4), 13-27.
38. Zhang, H. B., Zhang, Y. X., Zhong, B., Lei, Q., Yang, L., Du, J. X., & Chen, D. S. (2019). A comprehensive survey of vision-based human action recognition methods. *Sensors*, 19(5), 1005.
39. Zhang, J., Lin, H., Nie, W., Chaisorn, L., Wong, Y., & Kankanhalli, M. S. (2015). Human action recognition bases on local action attributes. *Journal of Electrical Engineering and Technology*, 10(3), 1264-1274.
40. Zhu, F., Shao, L., Xie, J., & Fang, Y. (2016). From handcrafted to learned representations for human action recognition: A survey. *Image and Vision Computing*, 55, 42-52.
41. Zelnik-Manor, L., & Irani, M. (2001, December). Event-based analysis of video. *In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001* (Vol. 2, pp. II-II). *IEEE.*