e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 6, Issue 6, June 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.54

International Journal Of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54| Monthly Peer Reviewed Journal|

| Volume 6, Issue 6, June 2023 |

A Comprehensive Description of The Machine Learning Processes For Human Action Realisation

Jiya Ram [1], Sumiran[2], Sumit Dalal[3]

P.G. Student, Department of ECE, Sat Kabir Institute of Technology and Management, Haryana, India [1]

Assistant Professor, Department of ECE, Sat Kabir Institute of Technology and Management, Haryana, India [2,3]

ABSTRACT: Systems for recognising and interpreting human behaviours properly employ data gathered from a variety of sensors. Automated and accurate recognition of human actions is one of the most difficult problems for computer vision. Due to the broad adoption of deep learning (DL)- enabled attributes, there has been a considerable rise in feature learning-enabled depictions for recognising actions in the past few years. This paper examines recent advancements in computer vision (CV) and gives a thorough analysis of human activity recognition. Security, home monitoring, augmented reality, human-computer interaction, and surveillance cameras are a few examples of computer vision applications that frequently work in tandem with human action detection. We provide a thorough, taxonomy-based analysis of human activity recognition (HAR) methods.

KEYWORDS: Human Action Recognition (HAR), Deep learning, computer Vision

I. INTRODUCTION

In recent years, the (HAR) has drawn a lot of interest since it can be used in multimedia surveillance equipment. Due to the wide range of human actions in everyday life, HAR is a difficult endeavor [1]. Owing to the sizeable video sequences that have to be analysed in surveillance systems, numerous (CV)-based solutions have been put out in the scientific literature but have not shown to be effective. When multi-view cameras are present, the issue gets worse. Deep learning (DL)-based algorithms have recently demonstrated great performance for HAR as well as multi-view systems with cameras [2]. considering its applicability in applications for video surveillance, HAR research is expanding rapidly [3]. In visual surveillance, HAR is crucial in identifying subjects' behaviours in public spaces. Additionally, these kinds of devices are helpful for monitoring in smart towns [4].

There are numerous sorts of human behaviour. There are two basic categories into which these behaviours can be divided: deliberate actions and involuntary actions [5]. Since it is laborious and prone to mistakes to manually recognise these activities in real time, numerous CV strategies have been published in the research [6] to help with this task. The majority of the solutions that have been offered are built on traditional methods including form characteristics, texture characteristics, point characteristics, and geometrical characteristics [7]. A few approaches [8] retrieve human silhouettes prior to the extraction of features, and some of them are based on the chronological data of the person [9].

II. RELATED WORK

Deep learning recently demonstrated promising outcomes in the area of computer vision (CV) [16]. By simulating the functioning of the human brain [17] to develop models, DL makes learning and data visualisation at many levels. Multiple processing layers, including convolutional, ReLu, pooling, fully connected, and Softmax, are used in these models [18]. The purpose of a CNN model is to simulate how the human brain maintains and interprets multidimensional data. The DL can be accomplished using a variety of techniques, such as neural networks (NN) that span several layers, hierarchical probabilistic theories, supervised learning, and unsupervised learning systems [19]. Due to the wide range of human behaviours that occur every day, the HAR process is a difficult undertaking. The DL models are employed to address this issue. The quantity of training samples is always a factor in how well a deep learning model performs. Multiple data sets for the action recognition challenges are accessible to the general audience. These datasets contain a variety of motions, including jogging, walking, getting out of a car, waving, punching, boxing, threw, slipping and falling, bending down, and a lot more.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54| Monthly Peer Reviewed Journal|

| Volume 6, Issue 6, June 2023 |

The majority of the most recent suggested systems concentrate on hybrid approaches, although they do not prioritise reducing computing time [21]. Given that most surveillance is carried out in real-time, this is a crucial consideration. Other significant HAR issues include the following: (i) The resolution of query video sequences is essential for identifying the focal point in its most recent frame. The multifaceted nature of the background, shadows, lighting, and clothing illnesses determine unimportant data from human action using traditional methods, leading to ineffective action categorization; (ii) it is challenging to identify proper human activities using automatic activities comprehension under multi-view cameras.

Changes during motion cause the improper activities to be captured by multi-view cameras; (iii) skewed datasets have an effect on a CNN's capacity to learn. A CNN model will always require a large number of training images to learn; also, (iv) the mining of characteristics from the complete video sequences contains a number of insignificant characteristics that have an impact on the precision of the classification.

III. FEATURE EXTRACTION-BASED ACTION IDENTIFICATION

For the identification of human actions, relevant motion data must be extracted from raw video data. The results of the HAR approach are directly impacted by the movement feature selection.The human body's look, the environment, and video cameras are only a few examples of numerous factors that might have a diverse impact on a single feature. As a result, action recognition's precision is constrained. Fig. 1 shows the feature extraction approach for HAR.

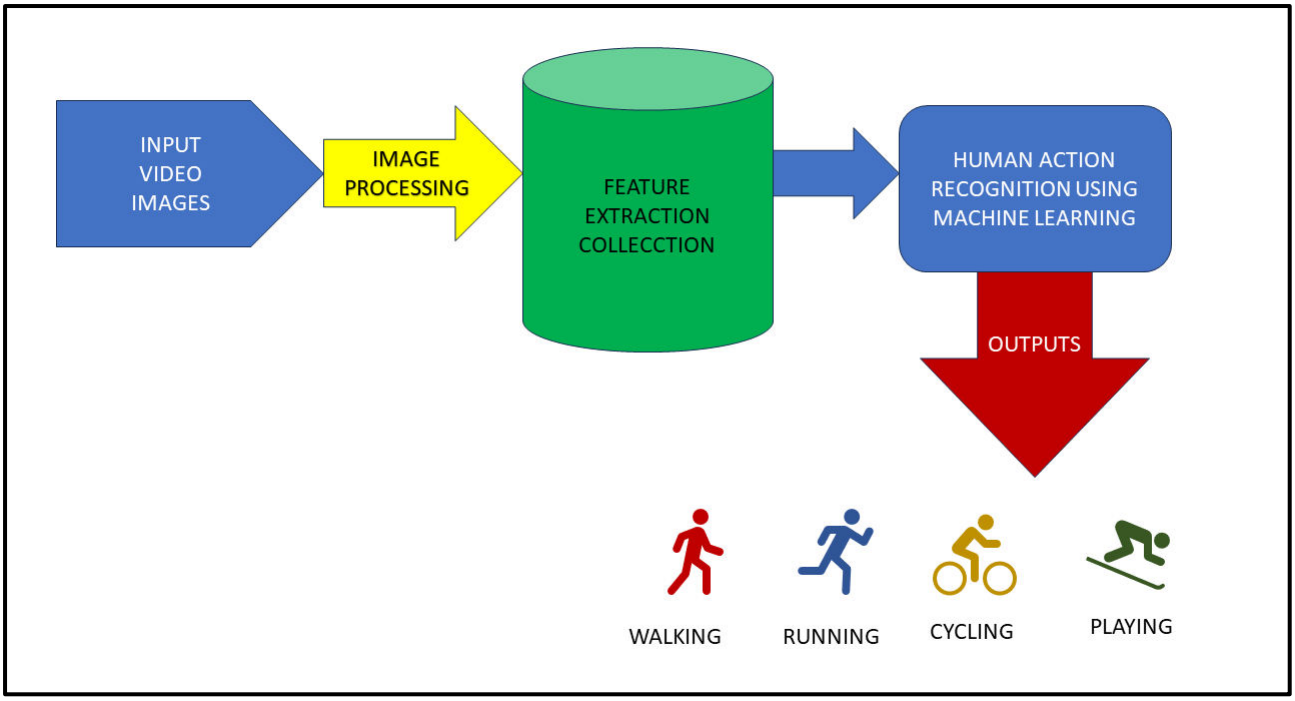

Fig. 1: Feature Extraction Based HAR

IV. HANDCRAFTED ILLUSTRATION METHOD

On a variety of action classification challenges, handcrafted visualisation of features has shown outstanding results. This method aims to extract local descriptors from video frames and recover the temporal and spatial information in videos. These characteristics, such as SVM and likelihood outline models, are frequently used in conventional ML to identify activity in unprocessed films. Perceptions of people and past context are used in handcrafted ways to draw practical conclusions from data. These types of techniques usually consist of three primary stages: (1) activity segmentation (2) choosing features (3) action categorization, depending on captured characteristics. Specific characteristics are retrieved from the original video sections and used to construct the descriptor. A general-purpose extractor is used to categorise the data, increasing the method's versatility, resulting in cheaper computational expenses,

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54| Monthly Peer Reviewed Journal|

| Volume 6, Issue 6, June 2023 |

and removing the need for massive training data sets. The handmade technique can be divided into three categories according to the type of data being used: techniques based on depth, strategies based on corpses, and approaches based on hybrid aspects.

The two primary groups of vision-based HAR techniques may be determined by a thorough analysis of the literature. (1) The standard manual approach based on handcrafted representations, which is based on expert developed detectors of features and descriptions including Hessian3D, (SIFT), (HOG), (ESURF), and (LBP). A general trainable predictor for HAR is then used, as demonstrated in Figure 2.

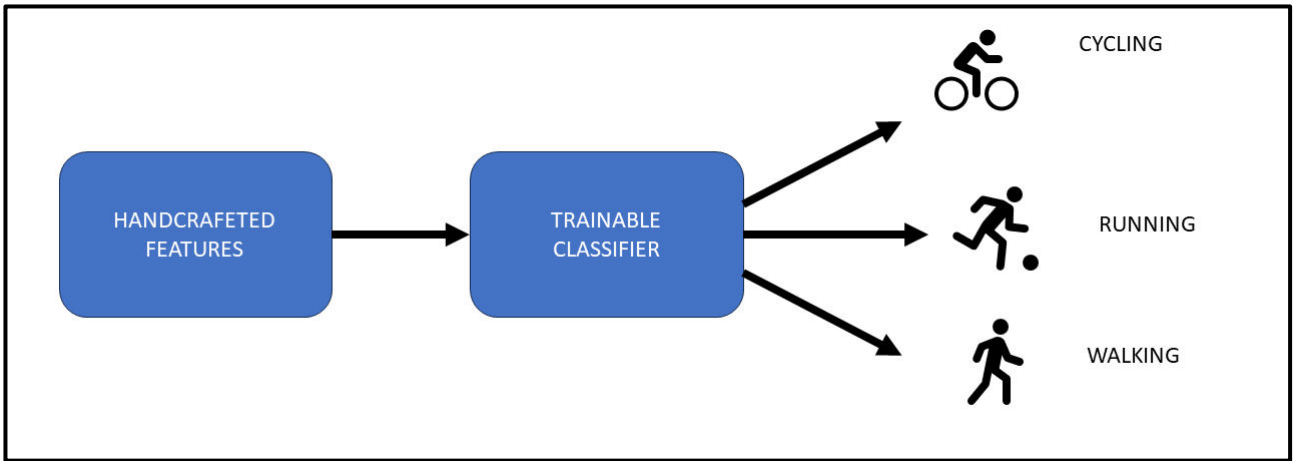

Fig 2.: Handcraft Feature Based HAR

The handcrafted representation based method largely adheres to the bottom-up HAR methodology. Foreground identification, manual feature extraction and portrayal, and classification are the three main stages. In the survey studies, many taxonomies have been utilised to discuss HAR methods. Though hierarchical techniques recognise more complicated actions by breaking them down into simpler activities (sub-events), single-layered approaches only recognise the simple activities from a video's sequence. In accordance with the feature description and method of classification employed for recognition, these are subsequently sub-categorized into space-time volumes and trajectories, presents a thorough overview of object segmentation techniques, covering the difficulties, tools, libraries, and free public datasets that can be used for object segmentation. The three levels of HAR, comprising basic technology, HAR systems, and applications, were covered in another study. Challenges like occlusion, anthropometry, execution rate, backdrop clutter, and camera motion have a substantial impact on activity identification systems [14].

V. DEPTH BASED APPROACHES

Scientists can now perform HAR with greater accuracy thanks to advancements in depth cameras and range methods of imaging [15]. RGB-D cameras collect depth data in addition to the standard RGB data to help computers recognise human activities more precisely. The human body and its associated movement in action are extracted via foreground detection in depth-based algorithms for HAR, which take the depth images as source. In order to identify the features, a number of investigators [16] reflected the depth data of a frame of an image in 3D, including views from the top, front, and side. In order to acquire the motion and shape characteristics for an aHAR model, the 3D points gathered from the surface image frame can be employed to compute the normal vectors. Using depth pictures, time, and coordinates, this descriptor may concurrently gather shape and motion data from a 4D space normal orientations histogram.

The segmentation of depth data to the point of concern and the extraction of characteristics for activity detection, alternatively, was introduced by a number of researchers. In order to minimise noise and obstruction, Wang et al. [16] introduced a technique for identifying semi-local features that explored a large sampling space. Authors in [17] provided a mechanism for displaying regional elements encircling the point of focus using videos. A local point descriptor to depict human activity in a depth environment was proposed in [18] and was acquired through collecting movement and structural information. Spatial-temporal interest points (STIPs) were created by Liu et al. [19] using structure and mobility characteristics that were recovered from noisy depth data. They developed a system of visual language, a two-tiered model that conveyed both shape and motion signals while removing noise. Yet, the computational expense and requirement to identify interest points using entire depth information from films limit the research potential of these systems.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54| Monthly Peer Reviewed Journal|

| Volume 6, Issue 6, June 2023 |

VI. DEEP LEARNING REPRESENTATION METHOD

A subset of machine learning called deep learning (DL) use hierarchical algorithms to extract high-level abstractions from data. It is a well-known strategy that has been widely applied in traditional AI fields like semantic parsing, transfer learning, NLP, CV, and many more. Two of the most important factors that led to DL enormous growth were the emergence of large, excellent, and freely accessible identified datasets and the availability of parallel GPU computing, permitting the transition from CPU-enabled to GPU-based training and, thus, enabling a significant acceleration in deep model training.

The DL family of methods includes NN, hierarchical probabilistic frameworks, and a number of unsupervised and supervised learned feature procedures. Because of their capacity to outperform advanced techniques in an assortment of tasks and the accessibility of different data from numerous sources, using DL have recently attracted attention. By switching from features that were manually generated to attributes based on DL, the CV has achieved notable results. The outstanding performance and strength of DL-based HAR in extraction attributes from multi-dimensional datasets is drawing a lot of attention these days. The DL techniques place each characteristic into a deep network as source and learn the complicated information over multiple layers, in contrast to the classical ML hand-crafted method where characteristics need to be modelled for HAR. Deep learning models require a lot of training data and are highly computational. These models are designed to learn multiple representational frameworks that provide automatic feature extraction for action recognition. CNNs, Autoencoders, RNN, and Hybrid Models are several deep learning-based action recognition methods.

VII. CONVOLUTIONAL NEURAL NETWORKS (CNN)

The (CNNs) are well known for being a pioneering DL technology with numerous robustly constructed layers. It has been demonstrated to be highly precise and is the one that is used the most frequently in various CV jobs. The overall CNN design is shown in Figure 4. Convolutional, pooling, and fully connected NN layers, respectively, are the three main types of NN layers that make up a CNN. Multiple layers serve various purposes. The network is trained using a forward process and a backward approach. The main goal of the forward process is to represent the input image with the present parameters (weights and bias) in each layer. After that, the losses are computed using expected and actual data.

VIII. POSE DETECTION AND LSTM (LONG SHORT-TERM MEMORY)

Using an off-the-shelf pose detection system to find the important points of a person's body for every shot of a video and then feeding those key points to an LSTM network to identify the activity being done in the video is another intriguing concept (Fig. 5).

IX. DEEP NEURAL NETWORK BASED

Deep neural networks find it tough and challenging to simulate human behaviours like body motions. Recognition of human activity is only recognition of human gesture. A gesture is the movement of bodily parts to express an important message. Let's say we have a lengthy film with numerous activities occurring at various points in the video. In these circumstances, we can employ a method known as Temporal Activity localisation (TAL). The model has a two-part architectural structure. Each particular activity is localised into temporal proposals in the first section. The second component then assigns a category to each video clip or proposal (Fig. 6).

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54| Monthly Peer Reviewed Journal|

| Volume 6, Issue 6, June 2023 |

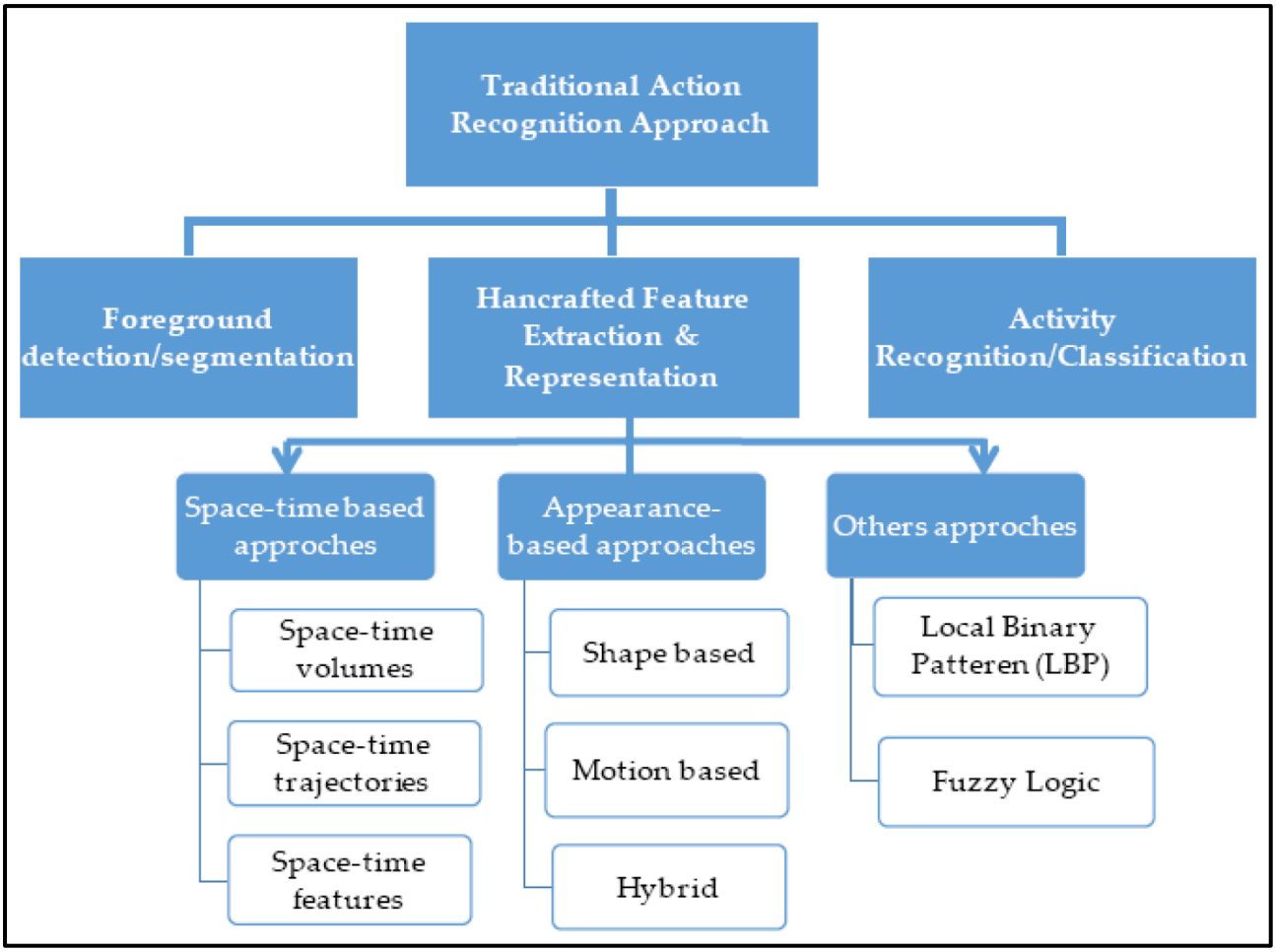

Fig.3: Conventional action illustration and recognition methods[14]

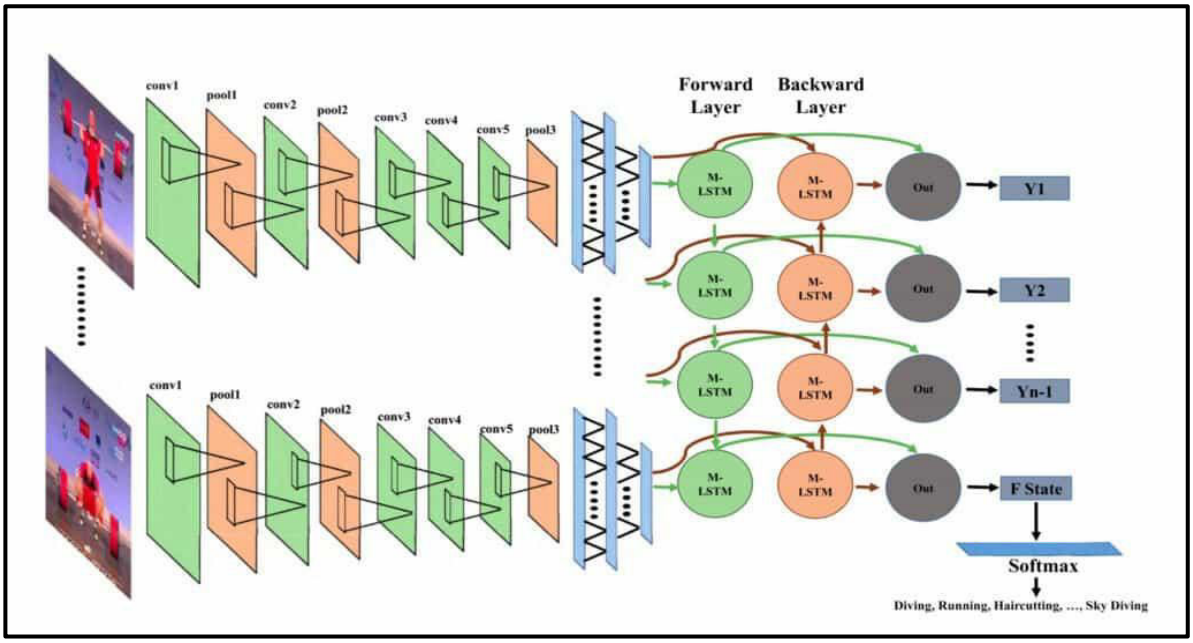

Fig. 4: CNN with LSTM Design [20].

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54| Monthly Peer Reviewed Journal|

| Volume 6, Issue 6, June 2023 |

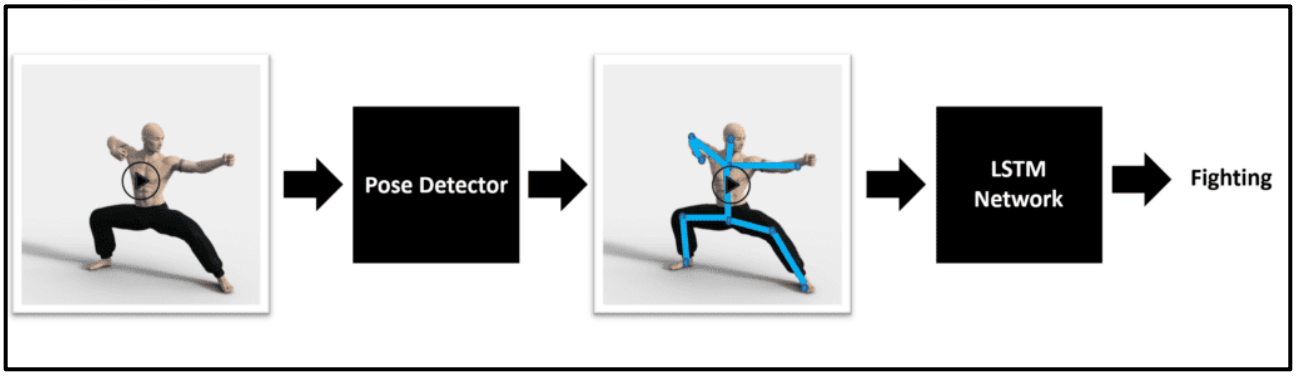

Fig. 5: LSTM based Posed Detection [20].

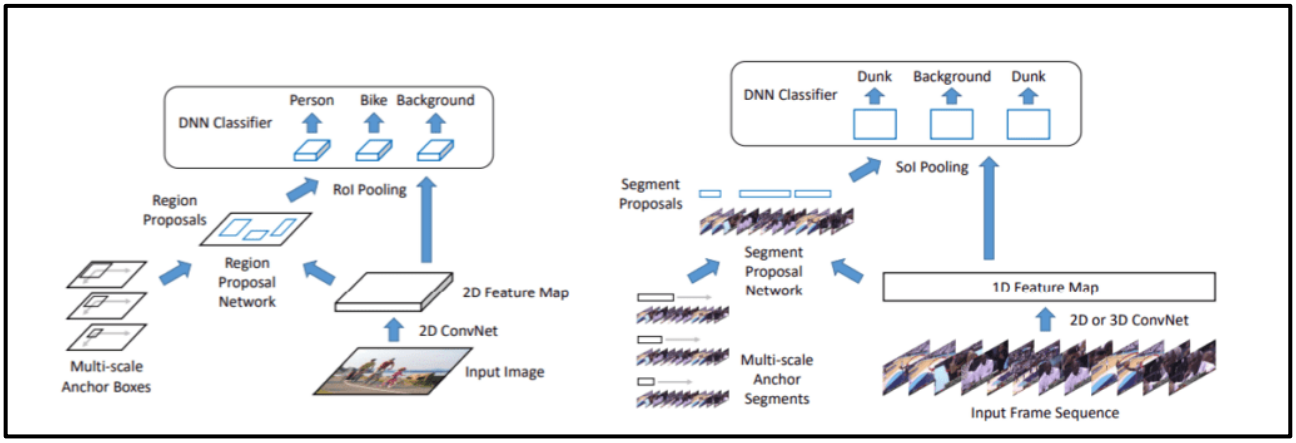

Fig. 6: TAL Design

X. CONCLUSION

In order to effectively understand and perceive human behaviours, computer vision is required in several fields, including human-computer interaction, robotics, monitoring, and security. This article offers a comprehensive overview of the most recent developments in this field of study. It outlines a number of standards by which it categorises things. This study starts by outlining the different HAR systems and their main objectives. After that, an overview of the techniques that are currently thought of as "state-of-the-art" was given, along with the validation techniques that were used to test those approaches. Additionally, it categorised human actions and the approaches applied to represent specific action data The different approaches can also be divided into groups based on the equipment used for data collection, and there is also a further division into detection, tracking, and recognition stages. Examining at the research's findings, it was discovered that every approach has its limitations. This may be attributable to advancements in the deep learning technique and encouraging results in terms of performance in identification and recognition. In contrast, group interactions and activities are crucial research topics because they may provide pertinent data in a range of HAR sectors, including public safety, camera monitoring, and the detection of anomalous behaviour.

REFERENCES

[1] Kim, D.; Lee, I.; Kim, D.; Lee, S. Action Recognition Using Close-Up of Maximum Activation and ETRI-Activity3D LivingLab Dataset. Sensors 2021, 21, 6774.
[2] Mishra, O.; Kavimandan, P.S.; Tripathi, M.; Kapoor, R.; Yadav, K. Human Action Recognition Using a New Hybrid Descriptor. In Advances in VLSI, Communication and Signal Processing; Springer: Singapore, 2021.
[3] Zin, T.T.; Htet, Y.; Akagi, Y.; Tamura, H.; Kondo, K.; Araki, S.; Chosa, E. Real-Time Action Recognition System for Elderly People Using Stereo Depth Camera. Sensors 2021, 21, 589
[4] Farnoosh, A.; Wang, Z.; Zhu, S.; Ostadabbas, S. A Bayesian Dynamical Approach for Human Action Recognition. Sensors 2021, 21, 5613.
[5] Hassaballah, M.; Hosny, K.M. Studies in Computational Intelligence. In Recent Advances In Computer Vision; Hassaballah, M., Hosny, K.M., Eds.; Springer: Berlin/Heidelberg, Germany, 2019.

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.54| Monthly Peer Reviewed Journal|

| Volume 6, Issue 6, June 2023 |

[6] Kolekar, M.H.; Dash, D.P. Hidden markov model based human activity recognition using shape and optical flow based features. In Proceedings of the 2016 IEEE Region 10 Conference (TENCON), Singapore, 22–25 November 2016.
[7] Hermansky, H. TRAP-TANDEM: Data-driven extraction of temporal features from speech. In Proceedings of the 2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No. 03EX721), St Thomas, VI, USA, 30 November–4 December 2003.
[8] Krzeszowski, T.; Przednowek, K.; Wiktorowicz, K.; Iskra, J. The Application of Multiview Human Body Tracking on the Example of Hurdle Clearance. In Sport Science Research and Technology Support; Cabri, J., Pezarat-Correia, P., Vilas-Boas, J., Eds.; Springer: Berlin/Heidelberg, Germany, 2016.
[9] Hassaballah, M.; Hosny, K.M. Studies in Computational Intelligence. In Recent Advances In Computer Vision; Hassaballah, M., Hosny, K.M., Eds.; Springer: Berlin/Heidelberg, Germany, 2019.
[10] Hassaballah, M.; Awad, A.I. Deep Learning In Computer Vision: Principles and Applications; CRC Press: Boca Raton, FL, USA, 2020.
[11] Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep learning for computer vision: A brief review. Comput. Intell. Neurosci. 2018
[12] Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent trends in deep learning based natural language processing. IEEE Comput. Intell. Mag. 2018, 13, 55–75.
[13] Palacio-Niño, J.-O.; Berzal, F. Evaluation metrics for unsupervised learning algorithms. arXiv 2019, arXiv:1905.05667.
[14] Sargano, A.B.; Angelov, P.; Habib, Z. A Comprehensive Review on Handcrafted and Learning-Based Action Representation Approaches for Human Activity Recognition. Appl. Sci. 2017, 7, 110.
[15] Nunez, J.C.; Cabido, R.; Pantrigo, J.J.; Montemayor, A.S.; Velez, J.F. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. Pattern Recognit. 2018, 76, 80–94.
[16] Chen, C.; Liu, K.; Kehtarnavaz, N. Real-time human action recognition based on depth motion maps. J. Real-Time Image Process. 2016, 12, 155–163.
[17] Wang, J.; Liu, Z.; Chorowski, J.; Chen, Z.; Wu, Y. Robust 3d action recognition with random occupancy patterns. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 872–885.
[18] Xia, L.; Aggarwal, J. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2834–2841.
[19] Liu, M.; Liu, H.; Chen, C. Robust 3D action recognition through sampling local appearances and global distributions. IEEE Trans. Multimed. 2017, 20, 1932–1947.
[20] https://learnopencv.com/introduction-to-video-classification-and-human-activity-recognition/

Impact Factor
7.54

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com