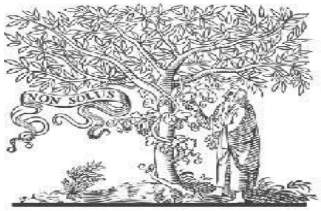
International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

COPY RIGHT



ELSEVIER SSRN

IJIEMR Transactions, online available on 05th Apr 2023. Link :http://www.ijiemr.org/downloads.php?vol=Volume-12&issue=Issue 04

10.48047/IJIEMR/V12/ISSUE 04/47

Title **DEEP LEARNING TECHNIQUES ON HUMAN ACTIVITY RECOGNITION FROM VIDEOS: A REVIEW**

Volume 12, ISSUE 04, Pages: 378-395
Paper Authors

Jeevan babu Maddala, Shaheda Akthar

USE THIS BARCODE TO ACCESS YOUR ONLINE PAPER

To Secure Your Paper As Per UGC Guidelines We Are Providing A Electronic Bar Code

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

Deep Learning Techniques on Human Activity Recognition from Videos: A Review

Jeevan babu Maddala[1]
[1]Department.of Computer Science and Engineering, Acharya Nagarjuna University, Guntur, India.
Shaheda Akthar[2]
[2]Department of Computer Science, Government College for Women (A), Guntur, India.
[1]jeevan.projects@gmail.com, [2]shahedaakthar76@gmail.com

Abstract

Over the last decade, there has been a fast growth of surveillance webcams in every part of human activity, resulting in a massive increase of camera footage. The main objective of this paper is to give an outline and comparative evaluation of new Deep Learning techniques for Human Activity Recognition (HAR) for different types of datasets. Our Novelty lies in the Evaluation and comparison of different Deep Neural Networks. HAR has several applications like Human Computer Interaction (HCI), Smart Driver Assistance Systems, Personal Assistant, Interactive games, Content based Video Annotation, Smart Medical Assistance systems, Smart Office systems, Smart Traffic Monitoring systems, Sports Analysis and Crime Scene Analysis. Following that, we examine and categorize current options offered to meet these objectives. All these applications face common set of challenges like noise, illumination changes, multi-view camera angles, partial occlusions, semantic classification of human activities. All the works are mostly based on Spatial and Temporal feature data for Activity recognition. Extraction of both temporal and spatial data from surveillance videos is required for successful video categorization. As a result, this work (CNN, RNN, LSTM, LSTM, and Multi-Stream Convnet Architecture) examines recent improvements in hierarchical conscience content deep learning architectures. It also dives into the various deep learning models available for HAR. We also go through the datasets (KTH, HMDB-51 and UCF-101), that were used for evaluation. This study seeks to provide the groundwork for future media HAR by identifying important challenges in the efficacy of human event detection in image sequences utilizing deep learning models. Finally, we present the accuracy comparison of different state of art techniques in HAR, followed by the disadvantages of current methodologies in the Deep learning approaches and conclude with futuristic trends in the human activity recognition.

Keywords: CNN, 3-D CNN, Deep Learning, Human Activity Recognition, LSTM, Multi-Stream Convnet, RNN, 2-Stream Convnet.

Introduction

Identifying people's behavior in footage is one of most promising applications of computer vision. This issue has recently sparked the interest of researchers from business, university, spy agencies, regulatory agencies, and the wider public. In the 1850s, contemporary photographers E. J. Marey and E. Muybridge undertook one of the earliest investigations into the nature of the human movement, photographing moving subjects and revealing various intriguing and artistic characteristics involved in humans and other animals locomotion. The classic Johansson [15] moving light display (MLD) experiment took the extensive investigations of motion capture awareness in the cognitive sciences forward significantly. The above paved the way for mathematical modelling of human action[43,44] as well as automated identification, all of which are inextricably linked to computer vision and pattern recognition.

A Human Activity Recognition (HAR)[26,27] may recognize individual activities and provide critical information to authorities so that they can take appropriate action [31]. Several instruments are present to record actions, including physical activity sensors, environmental sensors[50], heat sensors[48], and piezoelectric sensors [1,33], RADAR [17, 46], acoustic sensors [33,49], Echo, everyday things, and video recorders. Video-based HAR systems are frequently employed because to their numerous benefits.

A HAR method [32,37,45] can monitor individual activities and provide critical information to authorities, allowing them to take appropriate action [15]. Several sensors are present to record actions, including physical motion sensors, environmental sensors, thermal surveillance cameras, and piezoelectric sensors [8,31], RADAR [1, 17], microphones [46, 49], and webcams. Because of their multiple advantages, video based HAR systems are widely used. The HAR system's goal is to recognize and categorize real-world human activities. Human actions are complex and diverse, making effective activity identification difficult in computer vision.

Previous research on HAR systems has treated activity recognition as a standard pattern matching issue [46]. Support Vector Machine (SVM) and Hidden Markov Models (HMM) were used in early HAR approaches. Slowly, research has shifted to machine learning domain. In Traditional Machine Learning techniques, features are extracted manually with the help of domain experts, followed machine learning models. As a result, these models do not have the generalization property. Customized approximations are becoming obsolete as reinforcement learning provides direct extraction of features from video streams, obviating the requirement for domain specialists or optimal feature extraction.

In knowledge based methodologies, Artificial Neural Models are trained directly from data input such as pixels to classification. Deep learning techniques such as Deep Feed Forward Neural Networks and Feed Back Neural Networks are particularly successful in updating complicated actions due to their local dependency and scale invariance [49]. As a result, it's useful to gain a sense of current development in these disciplines,

as well as how the various methodologies compare. In recent years, human activity identification has been applied in a range of sectors.

HUMAN ACTIVITY RECOGNITION: APPLICATIONS

A. Smart Interactive Games:

User involvement is a critical component in the creation of an interactive game. The use of arm movements as the foundation for players to actively interact with game items shown on a flat plasma or LCD display has grown in popularity. It establishes a novel paradigm of interaction in which physical hand movements in the form of arm movements are coordinated with virtual items in the game. To engage with the games, the gamer employs quasi physical movements. For example, the well-known Microsoft Kinect Xbox or an interactive balloon game.

B. Smart Personnel Assistant:

Pose approximation and activity identification can also be used to help disabled persons, the aged, and regular citizens. For instance, a device that detects if a human collapses [33] or a blinking humanoids [2].

C. Autonomous Navigation Aided Systems:

A complete approach for intelligent driving assistance systems must include a focus only on the motorist [39]. A few more case studies of automated driving that use stance as well as behaviour investigations include observing approach that emphasizes depending on head pose monitoring [28], integrating motorist head pose and forearms monitoring for diversion warning system [37], predicting motorist leg movement to counteract brake conflations [38], developing intelligent airbag systems that rely on sitting body position evaluation [40], and forecasting motorist turn purpose [6].

D. Smart Human-Computer Interaction (HCI):

Despite traditional old type devices such as the computing keyboard and cursor pointing device, The main essential to design enhanced, more naturalistic medium between expert machines & users, with visual physical motion. Some Examples are Using arm movements to command slide presentation [15] or identifying industrial stages to enable people learn and improve their abilities.

E. Smart Physiotherapy:

Advanced biomechanics and physical therapy systems need the precise recording of healthy and malignant movement patterns, free of the artefacts caused by invasive marker-based camcorders. As a result, techniques for marker-less stance prediction and motion interpretation were proposed for use in this domain [41].

F. Smart Sports biomechanics:

Several sports, such as golf, cricket, and skating, need proper body movement patterns; thus, posture assessment [29] and gesture evaluation might be used to this sector for effectiveness and training analysis [45].

G. Automatic Video surveillance:

Smart Video surveillance is employed in a variety of situations, including essential services, trains and buses, commercial complexes, parking areas, and private residences. Continuously watching these cams, on the other hand, has become a concern. As a result, systems for

intelligent surveillance systems that include outside human activity analytics, such as [30], [42], will be required.

H. Intelligent Video annotation:

A massive number of visual data may be readily preserved with the advancement of industrial automation. There are many human-related movies among them, such as security footage, sports clips, and films. Instead of physically searching through massive video collections for the relevant paperwork, human gait interpretation can be utilised to interpret the image frames, for instance, ways of commenting football footage match [3] or, more broadly, approaches to annotate video of outside live sporting events [20].

HUMAN ACTIVITY RECOGNITION:

CHALLENGES

A. Real-World Conditions:

The majority of Human activity recognition systems are presently built and implemented on video frame taken under restricted settings. Distortion, occlusions, shadows, and other factors might significantly reduce the relevance in real - life settings. Mistakes in extracting features have a significant probability of propagating to greater levels. Human activity recognition algorithms [50,51,52] must be evaluated against such natural settings before they can be deployed in the field. Another important factor is the dealing with the high resolution and low-resolution images, which have significant impact on the accuracy of the HAR.

B. **Intra-Class Variances due to Camera Viewing***:* Finding approaches that can explain and are resilient to the huge variety in variables found within the same action class is one of the most critical issues in action recognition. While it is straightforward to develop statistical models of basic behaviors from a single perspective, extrapolating them to various perspectives is tremendously difficult. This is because webcam viewpoint effects and partial occlusion cause huge differences in motion and structural properties.

C. Intra-Class Variances due to Computation Rate:

The variance in execution speeds while completing the same action is the second primary source of observed variability in characteristics. In both inter-person and intra-person scenarios, there are differences in execution style. Because state–space approaches may not directly represent periodic axis changes, they are not really rate invariant and are subject to minor differences in execution rates.

D. Intra-Class Variances due to Biological changes:

Anthropometric variances, such as those caused by human shape, appearance, sexuality, and other characteristics, are another important class of variables that must be carefully considered. Unlike viewpoint and implementation variances, which have received considerable attention, morphological variations have just recently received a lot of attention. More studies need to be done assess the influence of anthropometric alterations but also to create approaches for achieving anthropometric invariance.

E. Sensors Assimilation:

A Computer vision device for recognising people's behaviour is regarded for important initial step in the development of machine intelligence systems. Connectivity the additional dimensions, including as sound, heat, movement, and proximity sensor, must be researched more thoroughly for the longer term goal of building powerful computational methods, or for the near - term goal of strengthening the reliability of activity detection algorithm.

DEEP LEARNING TECHNIQUES FOR HUMAN ACTIVITY RECOGNITION (HAR)

A. Overview:

The point of the study would be to investigate convolutional neural networks for identifying people's behavior. We divided deep neural architecture into eight categories. This study seeks to give scholars with resources they need to have a greater knowledge among the most recent approaches for video-based human activity detection that are being developed.

1) Convolutional Neural Network (CNN):

A Convolutional Neural Network (ConvNet/CNN) is a Neural Learning model that could accept an input image, assign relevance (learnable weights & biases) to various facets in the picture (.jpg), and distinguish good from bad. A ConvNet may over the Space - Time links in an image by using appropriate filters. The architecture delivers good adapting to the image dataset minimizing the number of input variables and the reuse of weights parameters. Convolution Layer, Activation Layers, Pooling Layer, and Classification Layers are the main layers of CNN.

2) Recurrent Neural Network (RNN):

It is a type of neural model intended to work with temporal data including sequences. Typical deep neural models are solely suited for unconnected variables. Furthermore, when we have information in a series in which one data set is reliant on the previous data set, we may modify the neural model to compensate for these correlations. RNNs have a concept known as "memory", which helps to maintain the information of prior inputs in construct of next outcome in the sequence.

3) Long Short-term Memory (LSTM):

Recurrent Neural Networks struggle with short-term memory. If the data series is large enough, they can have problems similar output across previous time steps to older books. As a result, if you're forecast anything from a large sentence of writing, RNNs may skip away critical data just at outset.

LSTMs and GRUs were created to fix the problem of short-term memory. They have internal mechanisms called gates that enable them to govern the flow of information. These gates can figure to see which input in a sequence should be saved and which should be deleted. This enables it to communicate critical details down the long network of patterns in order to make forcasting. These two networks provide nearly all state-of-the-art recurrent neural network results.You may even use them to make video captioning.

4) 3-D Convolutional Neural Network:

A 3D CNN is just the 3D equivalent: it receives a 3D space-time or a stream of 2D frames as data. Three-dimensional CNNs are an effective approach for understanding volumetric data embeddings.

5) **A Two-Stream Convnet Architecture:**

ConvNet architectures with two streams include both space - time neural models. Spatial CNN processes a stack of frames, whereas temporal CNN processes a stack of dense optical flow, each of which represents a particular video from the process database. This overall set up designed to improve the overall accuracy in the context of video event detection.

6) **Multi-Stream Convnet Architecture:**

Multi-Stream CNN refers to an architecture with far more than two streams. This deployment of a multi-input architecture that combines information from many perspectives of the same target in various aspects; hence, the extended multi-view design of Multi-stream-CNN allows it to make full use of limited picture data to increase recognition performance.

B. ***CNN Architecture for HAR:***

Moving Foreground Attention (MFA) is a revolutionary neural structure presented by Zhang et al. [51] that increases action detection accuracy by instructing the system to concentrate on discriminative objects. The shifting foreground is detected by MFA using a suggested variance-based approach. Simultaneously, an unregulated proposal is being utilised to explore activity core quantities and generate similarity scores. The MFA is trained using a newly devised stochastic-out technique based on these results. Experiments on UCF101 and HMDB51 were conducted out to establish their superiority over the other peer algorithms.

MC-SPRT[9] technique to improve accuracy of people activity recognition, although a random strategy is used to accelerate meaningful understandable video analysis. Based on the multi-class statistical tests, the consecutive stochastic ratio test is used to determine if a brief video can be classified into an activity labels. To prevent the tremendous processing price of video content understanding, the sampling method is halted again when the likelihood is high threshold. Based on an analysis of the UCF101 dataset, their findings indicate the usefulness of dynamic sampling approach.

Convolutional neural networks (CNNs)[52] be used to record spatial presentations and a linear dynamical system (LDS) be used to simulate motion information. They used image-trained CNNs to recognise motion clip ideas, which uses distinct levels of information by mixing the two layers in CNNs learned from pictures. The author [52] modelled the links between these clip notions using a linear dynamical system (LDS), which reflects the temporal patterns of activities. Finally, the authors stated that they used the suggested technique to two demanding actual benchmark functions, YouTube and UCF50, and obtained higher efficiency of up to 86.16 percent also on YouTube dataset and 82.76 percent also on UCF50 dataset.

A New Human Activity identification approach [19] based on deep learning is proposed in which every image in the live stream was tracked and identified as a human body. They then employed human silhouettes to generate binary spatial-temporal maps (BSTMs), in turn describe people activities over certain temporal frame. Eventually, they extracted characteristics from BSTMs and classified the actions using a convolutional neural network

(CNN). They assessed three publicly available data bases: Keck Gesture, KTH Database and Weizmann. Their metrics, such as recognition accuracy, are equivalent to modern deep learning approaches.

C. Recurrent Networks Architecture for HAR:

ATwo-Level Attention-Based Interaction Model [23] focused on multiple time variant cognitive processes to examine these colourful interactions. Personalized attention, which is influenced by postural attributes, employs various levels of interaction among persons inside a picture while refreshing actual values at each sampling interval. Their image level attention process makes use of an interest based pooling approach to study the various degrees of interconnections across people's activities and sematic level activity. Their approach comprises of a reformed Gated Recurrent Units (GRUs) network to control long-term temporal variability and consistency. Their entire trainable algorithm takes as input a succession of human detections in recordings or image clips and forecasts multi-person activity categories. To illustrate their efficacy, the author ran experiments on a volleyball dataset.

Furthermore, from the obtained attributes, a GRU framework was used to understand the periodic fluctuations in social behaviors. The proposed DGCF significantly lowered the consequences of comparable local movements and collected valuable background information by taking sub-groups into account. On the BEHAVE dataset, the proposed DGCF outperformed current framework approaches by 4.99 percent in prediction performance. Mostly New Collective Activity database, it also surpassed other techniques in classification accuracy by 3.75 percent.

Investigations is carried out to solve Network delay convergence problem in Z. Zhu [54] for training CNN parameters. As a result, rapidly training an effective CNN for action recognition is a difficult task. To ease network training, the author suggested an unique encoding termed nonlinear gated channels unit (NGCU) that encodes universal multi-channel interaction. Depending upon it, a nonlinear gated channels network (NGCN) for final encryption is constructed, and its final optimum performance is evaluated using the established benchmarks UCF101 and HMDB51.

D. 3-D CNN Networks Architecture for HAR:

ANovel Deep Learning Architecture, FSTCN [35], Sequential deep learning model, that acquires optimal spatio-temporal features via learning with standard Back-Propagation Error algorithms. This modular solution solves the dual problem of rising kernel overhead and a scarcity of learning video dataset. The T-P operator extends operations with extra capabilities & a temporal expression. Furthermore, the TCL uses two simultaneous kernels to learn more representative temporal aspects. They conducted extensive tests on the action benchmark datasets to demonstrate our algorithm's superiority.

The main issues of expensive 3-D network pre-training is solved in [13], by evaluating if the prior variables of 2D Deep networks could be quickly modularized into 3D. To address the training challenge, the authors devised a 2D-Inflated procedure as well as a concurrent 3D ConvNet structure. The 2D-Inflated approach transforms

pre-trained 2D ConvNets into 3D ConvNets without the need of webcam input. They explored the optimal number of 3D ConvNets in a parallel design and discovered that a 6-nets topology is an effective option for identification.

Residual attention unit (RAU)[24]is proposed to handle static distant pixels in feature maps by manipulating the data linked with the forefront zone. RAU is made up of spatial concentration and channel-specific awareness. The spatial awareness generates the attention mask by working on the intermediate feature map in a bottom-up top-down way, whereas the medium-wise awareness automatically reassesses the characteristic outputs of all streams. In addition, authors created a shortcut to retain the functionality of actual feature values by creating a quick link between the attention module's source and destination. Their suggested RAU is simply embeddable into 3D CNNs and allows for final learning alongside the models. Finally, tests on UCF101 and HMDB51 datasets were carried out to establish the validity.

The Hierarchical Multi-scale Attention Network (HM-AN) [47] is proposed in which combines the HM-RNN with the awareness mechanism and is utilised for activity identification. Gumbel-softmax, a recently proposed gradient approximation approach for probabilistic neuron, is utilized in the development of periodic border detector and the probabilistic strong attention approach. The authors conducted research on action identification from films to validate the efficiency of HM-AN. Experimentation indicated that their HM-AN outperforms LSTMs with attention.

E. Multi-Stream Networks Architecture for HAR:

A unique approach is introduced [7] for removing global camera motion dubbed Saliency-Context two-stream ConvNets, which directly generates similarity score over two successive images with no need for human observation. The proposed context two-stream ConvNets identify the complete context in image sequences, while prominent channel are tuned on prominent gait analysis motion areas observed in the warped optical flow. Finally, Saliency-Context multiple ConvNets enable us to collect complementary information while achieving cutting-edge performance on the UCF101 dataset.

A unique approach [16] for dynamically aggregating images in video stream for the job of recognising human behavior. In a single temporally search of such image sequence, their technique develops to aggregate such discriminative and relevant frames, while eliminating the bulk of irrelevant frames. The programme accomplishes this by continually estimating the exclusionary significance of each image sequence and then aggregating them in a deep network model.

RGB Rank Pooling Dynamic Network (RGB-RPDN) [53], a CNN architecture that maps a stream to several image-level dynamic domains of the same size as the input. They investigated how well the hand-crafted rank-pooling machine can represent stream development, and they employed Flow Rank Pooling Dynamic Network to extend the dynamic domain of the frame-level to that of the flow-level (Flow-RPDN). Experiment findings indicate that RPDNs outperform existing state techniques.

A Spatiotemporal Distilled Dense-Connectivity Network [11] is created for video action detection (STDDCN). STDDCN investigates interaction strategies between appearance and moving streams across topologies using these cognitive extraction and dense-networks. The availability of block-level tight linkages between visual and motion channels, in particular, enables space-time based interaction at the feature encoding levels. STDDCN's one-of-a-kind architecture allows it to progressively develop better cascading spatiotemporal properties. Finally, several ablation experiments on two benchmark datasets, UCF101 and HMDB51, validate the applicability & universality of our technique.

RESULTS AND DISCUSSION

Evaluating the machine learning algorithm is the most critical component of Human Activity Recognition. A conventional train/test/validation data base is to breakdown is to use 60% of the data for learning, 20% of the data for validation, and 20% of the data for verification. We create a third collection of data known as the validation set to check the model reliability.

When making classification predictions, there are following four outcomes.

- True Positives (TP) arise when an event is expected to correspond toward a single class and therefore is found to be a member of that class.
- True Negatives (TN) happen whenever we predict how an event will not correspond to just a label but it does not belong to that category.
- False Positives (FP) occur if we mistakenly believe that an event corresponds to a category whereas it's doesn't.
- False Negatives (FN) occur if we believe an event doesn't really pertain to a category when something actually it does..

A. Accuracy:

Accuracy is measured as the percentage of correctly predicted given the test suite. It's indeed easy to calculate by ration of the number of correct forecasts to the total number of estimates.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \text{---(1)}$$

B. Precision

Precision is described as the fraction of relevant evidence (true positives) across all samples predicted to pertain to a given category.

$$\text{Precision} = \frac{TP}{TP+FP} \text{------(2)}$$

C. Recall

Recall is characterized as the measure of instances expected to correspond to a class divided by the total number of examples that truly belongs to the group.

$$\text{Recall} = \frac{TP}{TP+FN} \text{------(3)}$$

D. *Mean Absolute Error*

The Mean Absolute Error is the mean of the variances between the Actual Values and the estimated Values. It tells us how far the projections differed from the actual outcome. Mathematically, it is represented as :

$$MeanAbsoluteError = \frac{1}{N}\sum_{j=1}^{N}\left|(y_j - \hat{y}_j)\right| \text{---(4)}$$

E. Mean Squared Error

Mean Squared Error (MSE) is identical to Mean Absolute Error (MAE), except MSE considers the mean of the square of the difference between the actual and projected measurements. MSE has advantage of making the gradient easy to compute, whereas Mean Absolute Error requires the application of complicated linear programming methods to evaluate the gradient.

$$MeanSquareError = \frac{1}{N}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2 \quad ----(5)$$

F. HAR: Datasets

Under this part, we discuss some of prominent or standard video clips used in the Literature of Human Activity Recognition. We list some of them below:

The KTH [34] dataset includes films of individuals executing six different sorts of actions: punching, hand-appreciation, hand-gesture, walking, marathon, and wandering. These activities are executed by 25 people in four scenarios: outside, outside with event size, outside under different clothes, and indoors. As a consequence, there are a total of 600 videos: 25x4x6 = 600. The images have a quality of 160x120 and a refresh rate of 25fps.

The Weizmann Human Action Dataset [10] is a freely donated dataset that contains 90 low quality clips (180144) of 9 different persons acting 10 different activities: sprinting, hopping , forward leaping, forward flexing, one hand flapping, leaping jack, lateral leaping, one-leg leaping, strolling, and Two-hand flapping.

UCF101 [36] is a data collection of 101 action categories of realistic action clips obtained from U tube. This set of data complements the UCF50 collection of data, that includes 50 activity categories. With 13320 videos from 101 action categories, UCF101 contains the highest diversity of actions, and is the highest complicated video clip set to present, with significant variances in camera movements, item appearance and position, entity scale, viewpoint, congested backdrop, illumination circumstances, and so on.

HMDB-51 [22] is indeed a gait analysis recognition video clips that includes 51 activity classes and over 7,000 physically cleared clips extracted from various sources ranging from digital motion films to YouTube. The dataset comprises 51 distinct activity classes, each with at least 101 clips, for a maximum of 6,766 video cuts culled from various input application domain. Each clip's label includes the wide - angle lens, image resolution, and amount of actors involved in the action.

Kinetics 700 [5] is a source of huge, high-resolution video data that include URL links close to 650,000 short clips, each of which covers 400/600/700 human activity categories, relying on the data release. Human-object interactions, such as playing musical instruments, as well as human-human interactions, such as holding hands and caressing, are featured inside the movies. Each activity category has at least 400/600/700 youtube clips.

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

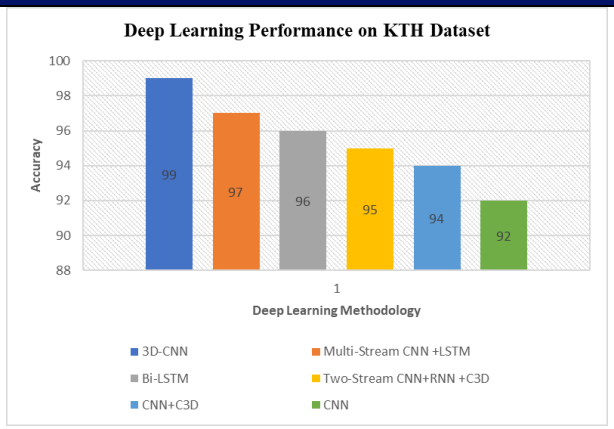


Fig.1. Description

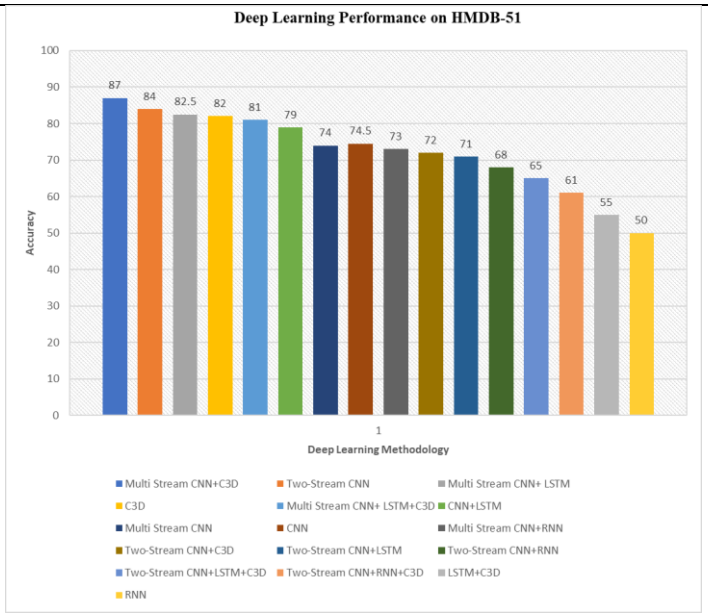


Fig.2. Description

International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

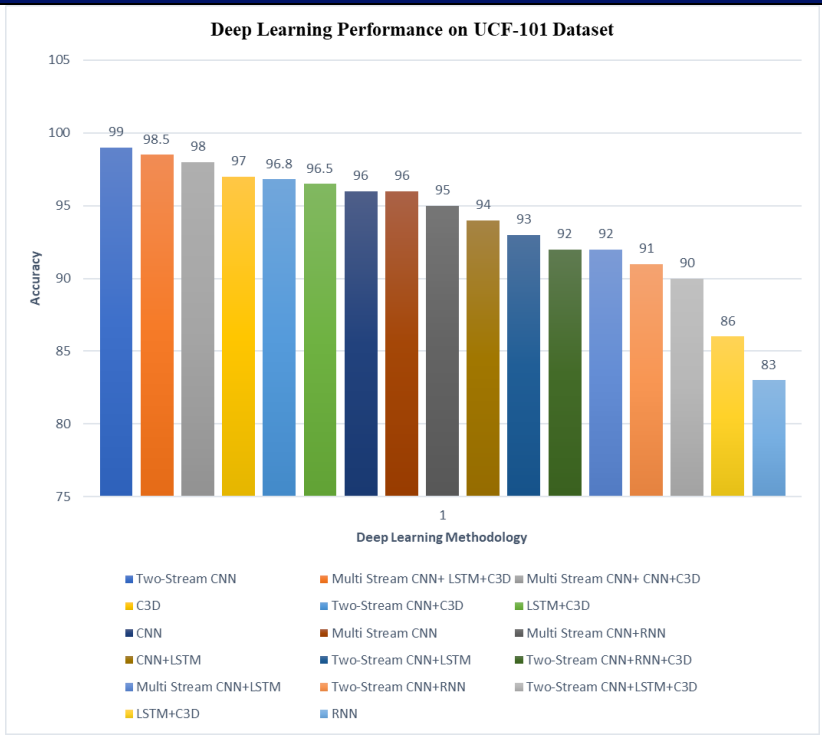


Fig.3. Description

Accuracy Comparison on Various Standard (a) KTH Dataset, (b) HMDB-51 Dataset and (c) UCF-101 Dataset

Every video footage is labelled by hand with a single action category and seems to last approximately ten seconds.

G. *Analysis*

This section emphasises the relevance of an diferent works that is appropriate for a certain data collection. Furthermore, first most significant HAR approaches in terms of effectiveness are fully explored. We thoroughly examine the various cutting-edge approaches and published datasets.

Figure 1(a) compares the top scoring neural architectural designs on KTH data collection. KTH dataset consists of 25 subject with various scenarios like inside & outside environment, scale variant scenarios. There are six different classes on a total of 2391 videos. The accuracy performance of 3D-CNN, Multi-stream CNN+ LSTM, Bi-LSTM

networks, Two Stream CNN+RNN+C3D, CNN+C3D and CNN on KTH data are 99%, 97%, 96%, 95%, 94%, 92% respectively. In all the above works, GoogLeNet, 3D DNN are used as feature extractors to perform classification. Any human action in a particular frame depends on its previous frames. Obviously, it is expected that Feedback networks like LSTM, RNN to perform better. But contrary to our belief, 3D Neural Networks performed better than other networks. After further analysis,the main reason is found out to be its low-level pixel feature representation for every action. Therefore, we can conclude that 3D-CNN or its extended version of 3D-CNN performs better for KTH datasets.

HMDB-51 dataset consists of different videos containing movie scenes, public databases and internet videos. They considered five human actions recorded in different view points and motion variations. In figure 1(b), we easily infer Multi-stream CNN+C3D performs better than Sequential networks like CNN, Two-Stream, Multi-Stream and feedback neural networks like LSTM, RNN. Investigations indicate that 2D + 3D multi-stream framework methodology based on valued patches surpasses earlier fusion approaches, achieving an accuracy of up to 87.7 on the HMDB-51 dataset [22], as shown in Fig. 1. (b).

UCF-101 dataset consists of web videos containing five human actions and recorded in unconstrained outside environments. 101 signifies the number of classes and with each class containing 100+ clips. In Figure 1(c), it is clear that Multi Stream/Two Stream with LSTM/C3D network perform with an accuracy of above 98% accuracy. All their methods divides films into overlapping frames before performing segmentation with localised sparse segmentation using global clustering (LSSGC), yielding the best results in comparison as shown in figure 1(c).

CONCLUSION

This paper presents a review and comparative assessment of current breakthroughs in Human Activity identification using multi-view data in this work. We give an outline of many industry verticals as well as the requirements for proper operation.

First, we discussed at some of the applications for detecting genuine human body activity using volumetric data. Smart Interactive Games, Personnel Assistant, driving assistance systems, Human-Computer Interaction, Smart Sports Biomechanics, Automatic Video Surveillance, Intelligent Video Annotation are some of the applications. We also spoke about some of the difficulties, such as Real-World Conditions, Camera Viewing, Computation Rate, and Biological Changes. Section IV summarises the brief description of the architecture, benefits, and limitations of Deep Learning approaches. The majority of these approaches are classified as follows: A. CNN, B. RNN, C. LSTM, D. LSTM, and E. Multi-Stream Convnet Architecture. Finally, we wrapped off the study with a review of several assessment criteria as well as accuracy comparison, often used in the field of Human Activity Recognition. Our conclusion is a combination of Multi-Stream and C3D networks performs in KTH, HMDB-51 and UCF-101 Dataset.

REFERENCES

[1] Al Hafiz Khan. M. A, Kukkapalli. R., Waradpande. P, Kulandaivel. S, Banerjee. N, Roy. N, and Robucci. R, "RAM: Radar-based activity monitor," in Proc. 35th Annu. IEEE Int. Conf. Comput. Commun. (INFOCOM), 2016; 1–9. https://doi: 10.1109/infocom.2016.7524361

[2] Alonso A., Rosa. R, Val. L, Jimenez. M, and Franco. S, "A robot controlled by blinking for ambient assisted living," in Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living. IWANN 2009. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg, 2009; 5518. https://doi: 10.1007/978-3-642-02481-8_127

[3] Assfalg. J et al., "Semantic annotation of soccer videos: Automatic highlights identification," Comput. Vis. Image Understand., 2003: 92(2): 285–305. https://doi.org/10.1016/j.cviu.2003.06.004

[4] Arshad, M.H.; Bilal, M.; Gani, A. Human Activity Recognition: Review, Taxonomy and Open Challenges. Sensors 2022; 22(17):6463. https://doi.org/10.3390/s22176463

[5] Carreira J, Noland E, Banki-Horvath A, Hillier C, and Zisserman A, "A short note about kinetics-600," 2018, arXiv:1808.01340. [Online]. Available: http://arxiv.org/abs/1808.01340

[6] Cheng S and Trivedi M, "Turn-intent analysis using body pose for intelligent driver assistance," IEEE Pervas. Comput., 2006:5(4): 28–37. https://doi: 10.1109/MPRV.2006.88

[7] Chen Q.Q, Liu F, Li X, Liu B.D., and Zhang Y.J, "Saliency-context two-stream convnets for action recognition," in Proc. IEEE Int. Conf. Image Process. (ICIP), 2016: 3076–3080. https://doi: 10.1109/icip.2016.7532925.

[8] Crasto. N, Weinzaepfel.P, Alahari. K, and Schmid. C, "MARS: Motion augmented RGB stream for action recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2019; pp. 7882–7891. https://doi: 10.1109/cvpr.2019.00807

[9] Fang H, Thiyagalingam J, Bessis N, and Edirisinghe E, "Fast and reliable human action recognition in video sequences by sequential analysis," in Proc. IEEE Int. Conf. Image Process. (ICIP), Sep. 2017: 3973–3977. https://doi: 10.1109/icip.2017.8297028

[10] Gorelick L, Blank M, Shechtman E, Irani M, and Basri R, "Actions as space-time shapes," IEEE Trans. Pattern Anal. Mach. Intell., 2007; 29(12): 2247–2253. https://doi: 10.1109/iccv.2005.28

[11] Hao W and Zhang Z, "Spatiotemporal distilled dense-connectivity network for video action recognition," Pattern Recognit., 2019; 92: 13–24. https://doi.org/10.1016/j.patcog.2019.03.005

[12] Hari Pavan, A., Anvitha, P., Prem Sai, A., Sunil, I., Maruthi, Y., Radhesyam, V. Human Action Recognition in Videos Using Deep Neural Network. Evolution in Signal Processing and Telecommunication Networks. Lecture Notes in Electrical Engineering, Springer, Singapore, 2022; 839.

https://doi.org/10.1007/978-981-16-8554-5_31

[13] Huang Y, Guo Y, and Gao C, "Efficient parallel inflated 3D convolution architecture for action recognition," IEEE Access, 2020; 8(1): 45753–45765. https://doi: 10.1109/access.2020.2978223

[14] Jain A, Gandhi. K, Ginoria D. K. and Karthikeyan. P, "Human Activity Recognition with Videos Using Deep Learning," 2021 International Conference on Forensics, Analytics, Big Data, Security (FABS), 2021; 1-5, doi: 10.1109/fabs52071.2021.9702599.

[15] Johansson. G, "Visual perception of biological motion and a model for its analysis", Perception Psychophys., 1973; 14(2): 201–211. https://doi.org/10.3758/bf03212378

[16] Kar A, Rai N, Sikka K, and Sharma G, "AdaScan: Adaptive scan pooling in deep convolutional neural networks for human action recognition in videos," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017:3376–3385. https://doi: 10.1109/cvpr.2017.604

[17] Khan. M. A. A. H, Hossain. H. M. S, and Roy. N, "Infrastructure-less occupancy detection and semantic localization in smart environments," in Proc. 12th EAI Int. Conf. Mobile Ubiquitous Syst., Comput., Netw. Services, 2015: 51–60. http://dx.doi.org/10.4108/eai.22-7-2015.2260062

[18] Khater, S., Hadhoud, M. &Fayek, M.B. A novel human activity recognition architecture: using residual inception ConvLSTM layer. J. Eng. Appl. Sci. 2022, 69(45), 1-16. https://doi.org/10.1186/s44147-022-00098-0

[19] Khelalef A, Ababsa F, and Benoudjit N, "An efficient human activity recognition technique based on deep learning," Pattern Recognit. Image Anal., 2019; 29(4): 702–715. https://doi.org/10.1134/s1054661819040084

[20] Kilner J, Guillemaut J.Y, and Hilton A, "3D action matching with key-pose detection," in Proc. Int. Conf. Comput. Vis. Workshops, 2009: 1–8. https://doi.org/10.1109/iccvw.2009.5457724

[21] Kim P.S, Lee D.G., and Lee S.W, "Discriminative context learning with gated recurrent unit for group activity recognition," Pattern Recognit., 2018; 76(1): 149–161. https://doi.org/10.1016/j.patcog.2017.10.037

[22] Kuehne H, Jhuang H, Garrote E, Poggio T, and Serre T, "HMDB: A large video database for human motion recognition," in Proc. Int. Conf. Comput. Vis., 2011; 2556–2563. https://doi: 10.1109/iccv.2011.6126543

[23] Lu L, Di H, Lu Y, Zhang L, and Wang S, "A two-level attention-based interaction model for multi-person activity recognition," Neurocomputing, 2018; 322(1): 195–205. https://doi.org/10.1016/j.neucom.2018.09.060

[24] Liao Z., Hu H, Zhang J, and Yin C, "Residual attention unit for action recognition," Comput. Vis. Image Understand., 2019: 189, Art. no. 102821. https://doi.org/10.1016/j.cviu.2019.102821

[25] Maha Mohammed Alhumyyani, Rasha Ismail, Mahmoud Mounir. Machine and Deep

Learning Approaches For Human Activity Recognition. International Journal of Intelligent Computing and Information Sciences (IJICIS), 2021;21(3):44-52. https://doi.org/10.21608/ijicis.2021.82008.1106

[26] Muhammad Ramzan, Adnan Abid, Shahid Mahmood Awan, "Automatic Unusual Activities Recognition Using Deep Learning", Computers, Materials & Continua, Academia, 2022; 70(1), 1829-1844, https://doi.org/10.32604/cmc.2022.017522

[27] Muhamada, A. W., & Mohammed, A. A, "Review on recent Computer Vision Methods for Human Action Recognition", ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, 2022; 10(4), 361–379. https://doi.org/10.14201/adcaij2021043613 79

[28] Murphy-Chutorian E and Trivedi M, "Head pose estimation and augmented reality tracking: An integrated system and evaluation for monitoring driver awareness," IEEE Trans. Intell. Transport. Syst., 2010; 11(2): 300–311. https://doi: 10.1109/tits.2010.2044241

[29] Murphy-Chutorian. E and Trivedi. M, "3D tracking and dynamic analysis of human head movements and attentional targets," in Proc. IEEE/ACM Int. Conf. Distrib. Smart Cameras, 2008: 1–8. https://doi: 10.1109/icdsc.2008.4635725

[30] Park S. and Trivedi M, "Understanding human interactions with track and body synergies (TBS) captured from multiple views," Comput. Vis. Image Understand., 2008: 111(1): 2–20. https://doi.org/10.1016/j.cviu.2007.10.005

[31] Ramamurthy S. R and Roy. N, "Recent trends in machine learning for human activity recognition—A survey," Wiley Interdiscipl. Rev., Data Mining Knowl. Discovery, 2018; 8(4): 1-16. http://dx.doi.org/10.1002/widm.1254

[32] Reinolds, F.; Neto, C.; Machado, J. (2022). Deep Learning for Activity Recognition in Real-Time Video Streams. Electronics. 2022; 11(5):782. https://doi.org/10.3390/electronics11050782

[33] Rougier. C, Meunier. J, St-Arnaud. A, and Rousseau. J, "Fall detection from human shape and motion history using video surveillance," in Proc. 21st Int. Conf. Adv. Inf. Netw. Applicat. Workshops, 2007: 875-880. https://doi: 10.1109/ainaw.2007.181

[34] Schuldt C, Laptev I, and Caputo B, "Recognizing human actions: A local SVM approach," in Proc. 17th Int. Conf. Pattern Recognit. (ICPR), 2004; 3: 32–36. https://doi.org/10.1109/icpr.2004.1334462

[35] Sun L, Jia K, Yeung D.Y, and Shi B. E, "Human action recognition using factorized spatio-temporal convolutional networks," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), 2015: 4597–4605. https://doi: 10.1109/iccv.2015.522

[36] Soomro K, Zamir A. R, and Shah M, "A dataset of 101 human action classes from videos in the wild," Center Res. Comput. Vis., 2012; 2(11): 2-7. https://doi.org/10.48550/arXiv.1212.0402

[37] Tran C and Trivedi M, "Driver assistance for 'keeping hands on the wheel and eyes on the road," in Proc. IEEE Int. Conf. Vehicular Electronics and Safety, 2009: 97–101. https://doi:10.1109/icves.2009.5400235

[38] Tran C, Doshi A, and Trivedi M, "Pedal errors prediction by driver foot gesture analysis: A vision-based inquiry," in Proc. IEEE Intell. Veh. Symp., 2011: 577–582. https://doi: 10.1109/IVS.2011.5940548

[39] Trivedi. M and Cheng. S, "Holistic sensing and active displays for intelligent driver support systems," IEEE Comput. Mag.,2007; 40(5): 60–68. https://doi: 10.1109/MC.2007.170

[40] Trivedi M, Cheng. S, Childers. E, and Krotosky. S, "Occupant posture analysis with stereo and thermal infrared video: Algorithms and experimental evaluation," IEEE Trans. Veh. Technol., Special Iss. In-Veh. Vis. Syst., 2004; 53(6): 1698–1712. https://doi: 10.1109/TVT.2004.835526

[41] Trivedi M, Huang K, and Mikic I, "Dynamic context capture and distributed video arrays for intelligent spaces," IEEE Trans. Syst., Man, Cybern., A, 2005: 35(1): 145–163. https://doi: 10.1109/TSMCA.2004.838480

[42] Utasi A and Benedek. C, "A 3-D marked point process model for multi-view people detection," in Proc. Computer Comput. Vis. Pattern Recognit., 2011: 3385–3392. https://doi: 10.1109/CVPR.2011.5995699

[43] Viet-Tuan Le, Kiet Tran-Trung, Vinh Truong Hoang, "A Comprehensive Review of Recent Deep Learning Techniques for Human Activity Recognition", Computational Intelligence and Neuroscience, 2022; Spl. Issue., 1-17 pages. https://doi.org/10.1155/2022/8323962

[44] Vijeta Sharma, Manjari Gupta, Anil Kumar Pandey, Deepti Mishra &Ajai Kumar (2022) A Review of Deep Learning-based Human Activity Recognition on Benchmark Video Datasets, Applied Artificial Intelligence, 2022; 36(1): 1-48. https://doi.org/10.1080/08839514.2022.2093705

[45] Waibel. A, et al., "Computers in the human interaction loop," in Handbook of Ambient Intelligence and Smart Environments. New York: Springer, 2010. https://doi: 10.1007/978-0-387-93808-0_40

[46] Wang. J, Chen. Y, Hao. S, Peng. X, and Hu. L, "Deep learning for sensor based activity recognition: A survey," Pattern Recognit. Lett., 2019: 119: 3–11. https://doi.org/10.48550/arXiv.1707.03502

[47] Wang J, Peng X, and Qiao Y, "Cascade multi-head attention networks for action recognition," Comput. Vis. Image Understand., 2020; 192, Art. no. 102898. https://doi.org/10.1016/j.cviu.2019.102898

[48] Xu, Y., &Qiu, T. T. Human Activity Recognition and Embedded Application Based on Convolutional Neural Network. Journal of Artificial Intelligence and Technology, 2020; 1(1):51–60. https://doi.org/10.37965/jait.2020.0051

[49] Yang J, Nguyen M. N, San. P. P, Li X, and Krishnaswamy.V, "Deep convolutional neural

networks on multichannel time series for human activity recognition," in Proc. IJCAI, Buenos Aires, Argentina, 2015;15: 3995–4001. https://doi.org/10.1109/ACCESS.2022.3192452

[50] Zhang, S.; Li, S.; Zhang, S.; Shahabi, F.; Xia, S.; Deng, Y.; Alshurafa, N. Deep Learning in Human Activity Recognition with Wearable Sensors: A Review on Advances. Sensors 2022; 22(4), 1476. https://doi.org/10.3390/s22041476

[51] Zhang J, Hu H, and Lu X, "Moving foreground-aware visual attention and key volume mining for human action recognition," ACM Trans. Multimedia Comput., Commun., Appl., 2019; 15(3): 1–16. https://doi.org/10.1145/3321511

[52] Zhang L, Feng Y, Xiang X, and Zhen X, "Realistic human action recognition: When CNNS meet LDS," in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), 2017: 1622–1626. https://doi: 10.1109/ICASSP.2017.7952431

[53] Zhu Z, Ji H, Zhang W, and Xu Y, "Rank pooling dynamic network: Learning end-to-end dynamic characteristic for action recognition," Neurocomputing, 2018; 317: 101–109. https://doi.org/10.1016/j.neucom.2018.08.018

[54] Zhu Z, Ji H, and Zhang W, "Nonlinear gated channels networks for action recognition," Neurocomputing, 2020; 386(1): 325–332. https://doi.org/10.1016/j.neucom.2019.12.077