Vision based human activity recognition using deep neural network

Kumari Priyanka Sinha[a], Alok Kumar[b,*] Ashish Ranjan[c]

[a]Department of Computer Science and Engineering, Nalanda College of Engineering, Chandi, Nalanda, India

[b,*]Department of Mechanical Engineering, Nalanda College of Engineering, Chandi, Nalanda, India.

[c]Department of Electrical and Electronics Engineering, Nalanda College of Engineering, Chandi, Nalanda, India

[*]corresponding author, email Id-alokbpsc2018@gmail.com

Abstract

Recently computer vision technology is continuously growing and the accuracy of the activity detection of human or object as well as target recognition has been improved. While working on human recognition face recognition is one of the important directions which can solve many challenges like mobile payment, safety, criminal investigation etc. Deep learning can play a vital role in the field of computer vision technology. So, in the proposed research study a neural computing model is introduced for human activity recognition and handling the other challenges. Some of the challenges like intra-class variations, image quality issues, illumination variations etc. in this research study new deep learning CNN classifiers is proposed. The experimental results shows that it achieves better result for human action recognition.

Keyword: *Activity recognition, HAR, deep leaning, CNN, vision*

DOI Number: 10.14704/nq.2022.20.11.NQ66323 **NeuroQuantology 2022; 20(11): 3117-3128**

Introduction

Nowadays we cannot imagine the world without videos and these have become an inseparable part of our lives. Although video cameras are used everywhere but especially in public places video surveillance cameras are used. With the growing video contents, it is easier to access them but the human capabilities are limited to analyse them. To analyse the huge video content there is a need for intelligent systems that could analyse the video content as well as recognize the human activities occurring in video. With fast and rapidly evolving computing resources such systems are possible which can perform activity recognition much faster than any human. Although there are ample research studies carried out for human activity recognition system, the growing technologies in the field and multi-disciplinary nature of human activity (Poppe, 2007) [01] recognition prompt the need for updates in the field. With the use of Human activity recognition intelligent system,the physical activitiescan be detected occurring in the video. HAR intelligent system aims to recognize the actions of a person automatically basedon video content and through machine learning methods.

Automatic human action recognition is useful for video surveillance and other human computer interaction applications. Different approaches are used to recognize human activities for example techniques for tracking are used which estimates body pose and others analyse pattern of appearance and motion from the video.Usually spatial-temporal operators and local descriptor techniques are used for human action representation.

The study of fundamentals of Digital Image Processing (DIP) gives a brief explanation of digital image processing. This is a wide research area which includes computer vision, video processing, image processing and human (or object) motion tracking and many. These are the fundamental learning for the action recognition venture. Digital Image Processing is an area of dealing with images in which the input and the output of the process

eISSN 1303-5150

are images. An image might be characterized as a two-dimensional function *f (x, y)*, where x and y are special (or plane) coordinates. The amplitude of *f* at any combination of two coordinates *(x, y)* is the intensity or Gray scale of the image at that point. Whenever *(x, y)* and amplitude of f are all finite and distinct quantities, then the image is referred as digital image.

Segmentation is a methodology, in which image is transformed into tiny segments such as it could be extracted as very accurate image attributes. Image segmentation separates an image into its component of regions or objects. This depends on features of an image such as point, line, edge and region. Thresholding plays an important role and is very much used in segmentation. It may be classified into two levels such as, single level and multi-level thresholding. If a single threshold value, for example T is used then it is called as single level thresholding. If a greater number of thresholding is used for separating the image information, then it is known as multi-level thresholding. The different types of thresholding are Global, Local and Adaptive thresholding. Global thresholding depends upon only the Gray levels of an image. Local thresholding depends upon Gray level and some local properties of the image. Adaptive thresholding also known as dynamic thresholding, that depends upon the Gray level, local properties and the spatial co-ordinates of the image pixels.

Representation and Description of image

Representation involves the steps that extract the attributes that are useful for processing through computer. After feature extraction of the image, raw data will result. To make them more suitable for computer processing the descriptors are used. Using Representation, a compact image can be attained before applying descriptors. Chain codes and Polygonal approximation are the most useful representation techniques.

Description is referred to as feature selection and it deals with extracting the attributes that produces some quantitative data of interest. It may also be used to differentiate one class of object from other. It can be

- Boundary descriptors
- Regional descriptors and
- Relational descriptors

Boundary descriptors are based on the external shape of feature in the image and regional descriptors are based on the internal properties of the image. But both descriptors do not provide the relation between different features, so the relational descriptors are used for that purpose.The approach used for recognizing human action from video is producing space-time shape in the spacetime volume, that is observed. Space-time shape consists of spatial information of pose of the human body and the body motion or the dynamic information. Another approach that can use the space-time shape of an action which uses motion history images to analyse planar slices of the space-time intensity volume(Blank, n.d.) [02].

Human action is represented as space time shapes and this shape is considered to be surrounded by simple closed surface. In this manner every internal space time point is to be assigned by their relative position in the space time shape. This is happened by assigning space time point as the mean time which is needed by a particle to undergo a random walk process. This random walk process starting from a point and hit the boundaries. It is computed by following equation-

$\Delta\, U(x, y, t) = -1$, and (x, y, t) ϛ $\in S$

Where Laplacian U is defined as $\Delta\, U = U_{xx} + U_{yy} + U_{tt}$ with the boundary conditions $U(x, y, t) = 0$, and the boundary condition is ∂S

eISSN 1303-5150

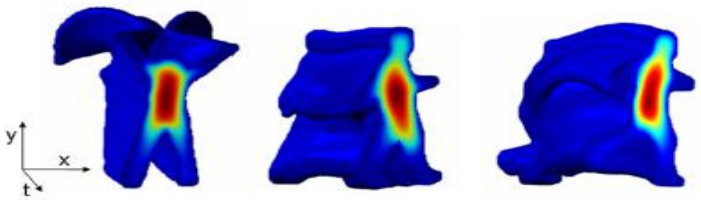

Figure 1: Human Action on Space Time Shape (Jumping, walking and running shapes)

Figure 1 represents spatial form to the Poisson equation for the space time shapes. A generalized analysis to identify the action in the form of space time shapes near any point is mentioned below. Let initially space time shape is represented by points (x, y,t) and satisfy the equation-

$$P(x,y,t) = ax^2 + by^2 + ct^2 + dxy + eyt + fxt + g \leq 0$$

And the Poisson solution is described in the form of-

$$U(x,y,t) = -\frac{P(x,y,t)}{2(a+b+c)}$$

While approaching to the centre, the value of U increases quadratically and U represents nested collection of conic boundaries. Now considering the Hessian matrix of U the matrix at given point can be obtained and it is mentioned as follow-

$$H(x,y,t) = -\frac{1}{a+b+c}\begin{pmatrix} a & d/2 & f/2 \\ d/2 & b & e/2 \\ f/2 & e/2 & c \end{pmatrix}$$

The above matrix represents the whole 3D conic shape which is scaled by a constant. The orientation of the shape and its aspect ratio is revealed by eigen values and eigen vectors of H.

Recognition

Recognition is a method that allocates a label to an object or image based on its descriptors. The main goal to classify data (image pattern or objects) is based on prior knowledge or on statistical information extracted from the pattern or object. Classification(Methods, 2019) [03] may consist of either supervised or unsupervised classification of the given pattern or object. Several approaches are available for object recognition or image recognition such as, neural networks, statistical classification, template, syntactic or structural matching.

Human action recognition includes the feature extraction from the given video frame and structural matching of those features with other video frames. The pattern or object recognition can be classified as supervised and unsupervised classification. In supervised classification, for example, in human action recognition the input image is classified as a number of predefined classes. In unsupervised learning, it is classified as unknown classes (for example clustering).

1. Background and Related Work

The protection of smart infrastructure such as smart bridges, overpasses, dams and tunnels from attacks is necessary. An smart video automated system (Bodor, n.d., 2006) [04] can detect the suspicious motion or activities near the critical infrastructure. The developed software recognizes the activity when they pass through the field of vision of camera by using vision algorithms to classify the motion and activities of humans. Most of the earlier research based on some novel methodologies for automatic tracking and pose estimation (Moeslund et al., 2006) [05] in natural scenes. Recent research trends focused on video based human capture and analysis to understand human actions and

eISSN 1303-5150

behaviour. The aim of the advanced research is to achieve automatic visual analysis of human motion activities.

The motivation of the research in this field is its various application areas like, surveillance, human computer interaction, and automatic analysis. To recognize human activities, human motion analysis and its characteristics are important. Human motion analysis can be done through modelling or construction of likelihood function and their estimation to find the most likely pose in the likelihood surface.Recent advancement in deep learning and availability (Zhao & Zheng, 2012) [06] of more powerful tools that can learn semantic, deeper features made it possible to recognize human activities. Through these advanced techniques, various tasks like salient object detection face detection and pedestrian detection can also possible. To improve the performance of human recognition techniques (Ye et al., 2015) [07] binocular vision methods which are based on binocular stereo vision and human face-hand feature can also be useful. These methods can solve the problem of occlusion and multi-direction movement of traditional video surveillance system. Video content analysis can be helpful in civil construction (Pereira et al., 2015) [08] for inspection of building pathologies and detecting of cracks in building facades. Theses video content is generated by Unmanned Aerial Vehicle (UAV) and the image or video processing algorithms can be embedded in UAV itself.Computer vision algorithms can be integrated with UAVs (Al-kaff et al., 2017) [09] to cope with certain difficulties in aerial perception like visual navigation, obstacle detection, and decision making. With the use of advanced techniques, the limitations like visual odometer, obstacle detection, localization and mapping etc.A real time vision based automated system (Abed & Rahman, 2017) [10] which can monitor objects is developed with camera module and it is programmed with Python Language and supported by Open Source Computer Vision (Open CV) library. Image processing algorithm monitors an object with the extracted features.

Human activity recognition framework(Roobini & Fenila Naomi, 2019) [11] observes the human movement using different deep learning approach. With the help of sensed data deep learning models can identify human motions with the high accuracy. The framework uses convolutional neural network with long short-term memory and recurrent neural network.In the proposed model(Vahora & Chauhan, 2019) [12] contextual relationship based deep neural network is used for activity recognition from video sequence. the proposed method relies on scene level feature extraction by using convolutional neural network. the proposed framework is very effective in recognizing group activities in video surveillance.Traditional methods of human activity recognition require huge data for training classifiers. These methods are not effective on weakly labelled data. In(Wang et al., 2019) [13] the proposed method the traditional methods are modified by attention mechanism to make compatible the global and local feature extraction. The proposed method relies on processing of sensor data and to collect it easier. It is also claimed to have less computational cost in compare to traditional CNN model. The proposed work can be extended for weakly labelled dataset, collected from sensors to improve the model to recognize and locate various activities.

Vision based techniques for human activity analysis(Golestani & Moghaddam, 2020) [14] is extensively used but it faces several challenges, viz. it requires infrastructural support such as installation of video cameras in each surveillance area which is prone to huge cost. Body sensors can be an

eISSN 1303-5150

option to collect and translate human motion into signal patterns for activity recognition.Recently Internet of Healthcare Things has become important for human activity recognition. it became possible due to rapid development of wearable and mobile devices. in (Zhou et al., 2020) [15] the proposed solution for human activity recognition an intelligent auto labelling scheme which is based on deep Q-network is developed to improve the learning efficiency in IoT environment.

2. Human Activity recognition (HAR)

This new approach of understanding human action in given video has two distinct scenarios, from simple action to complex video frames analysis. With this intention, first envelop the case of recognizing different actions, where the person in video is involved in one simple action.

Human activity recognition system collects the data from wearable sensors in the form of video frames or images and then process it for analysing the human information. This process involves extracting the features of raw data and employing these features to train or develop the techniques in HAR tasks. For action recognition vision-based data is collected and image sequences which are labelled with action tags are used. Figure 2 summarises the different techniques used in HAR. The HAR modelsemploy machine learning algorithms such as decision trees, SVMs, naive Bayes, and hidden Markov models (HMMs).These methods are widely used in HAR and similar researches.

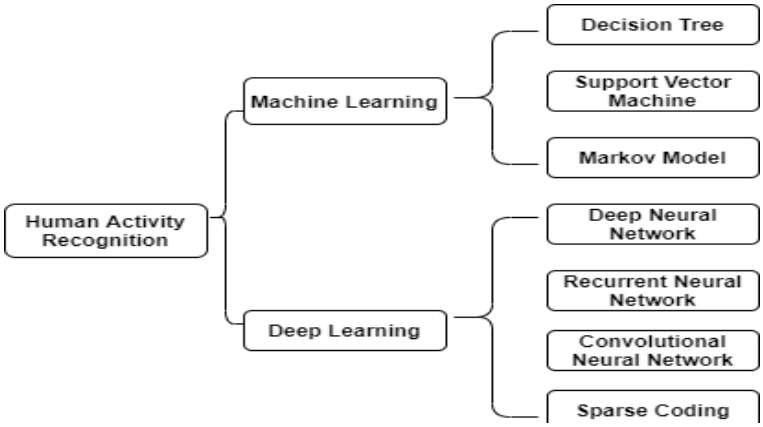

Figure 2. Different Methods for Human activity recognition

Theory

Feature Extraction

Image features play an important role in image processing. It is an interesting point in an image or a picture, and it is used as the initial point in many computer vision algorithms. Features are the main primitives for the feature detection algorithms. The feature detector algorithms are used to detect the feature values present in the image. In the proposed work, feature detection method was applied to extract the features of a person from the video frame. There are many feature extraction methods developed.

For features, X, Y and Z dimensions needs to be first merged and also include total acceleration magnitude (Mag). The total set of inputs now include: X, Y, Z, and Magnitude For every dimension, the features was extracted and compared in their importance. To evaluate the significance of the features, a leave-one-out approach is used in predicting accuracy for the full dataset. Also, a package

eISSN 1303-5150

specific implementation for highlighting the static importance of each feature is assessed. Test with leave-one-out was performed in groups of the following features left out:

- Frequency-based approach for all features
- Square of sums for percentiles in magnitude, for all features
- SC and ST individually held out
  For this, statistical features is used in both time- and frequency domain data of labelled sequences. Features include:
- Mean amplitude of window
- Variance of window
- Sum of difference between consecutive measurements
- Square of sum, 25th percentile in magnitude of feature
- Square of sum, 75th percentile in magnitude of feature
  All features are made from the time- and frequency domain signal of each dimension.

3. Deep Neural Network (DNN)

Deep neural network consists of input layer, two or more hidden layers and an output layer. Deep neural network has more hidden layers than traditional three-layer artificial neural network and hence has more parameters. DNN (Zhao & Zheng, 2012) [06] can learn classification function automatically from the sensor data with the help of large number of parameters. Human activity recognition is considered to be the time series classification problem.While using machine learning and deep learning models, LSTM (long short-term memory) model of Recurrent neural network (RNN) is used to recognize various activities of humans viz. standing, walking, sitting, climbing etc.

In human activity recognition while using LSTM RNN, the movement of human body can be classified in different categories like, walking, sitting, standing, laying, walking upstairs and walking downstairs. In recurrent neural network with LSTM, data can be directly fed into the neural network which acts as a black box. This method is not required feature engineering and comparatively simple in terms of classical data science techniques. RNN may takes many input vectors which helps in processing the images and produces the output in the form of other vectors. In this process time series of feature vectors is accepted to convert into probability vector as an output and for classification purpose.

DNN Process

1. Understanding the dataset includes pre-processing the data by applying noise filters and then sampled for readings.
2. While reading the data 80% of it taken as training data and remaining 20% of it taken for test data.
3. After uploading the dataset, a model is selected and build of deep learning with importing the necessary library.
4. While building model in training phase to achieve better accuracy LSTM model of RNN is also chosen.
5. Now pose vector is defined which can express human body in the form of location of all k body parts, and pose vector is defined as-

$$y = (\dots, y_i^T, \dots)^T, i \in \{1, \dots . k\}$$

Where yi is representing the x, y coordinates of the location of ith body joint. The whole image is represented in the form of (x, y) in which x is image data and y represent data of ground truth pose vector.

6. The pose vector then converted through image localization task using CNN architecture-

$$y^* = N^{-1}(\varphi(N(x); \theta))$$

Here $\theta$is trainable parameter, Ψ is neural architecture which is applied to the

eISSN 1303-5150

normalized pose vector N(x). y^* is the predicted output which is obtained by denormalization of output N^{-1}.

In human action recognition deep learning methods related to efficiently using CNNs in image classification(Mliki et al., 2020) [16]. The most advantageous thing of CNN is that they learn features as well as classification boundaries both. It is proved from many researches that CNN based approaches outperform for image classification and for action recognition also.

4. Proposed Model

Human action recognition is being tried from past many years, but still this field is immature. It also requires intense investigation to reflect various ideas and approaches. While performing human activity recognition it should be independent from the objects in the video frames since objects may also carrying some activities and their interactions. Here the entire human body need to be considered without focusing on specific human body parts. While using the holistic approach it would attempt to identify various information like gender, identity as well as simple actions like running, walking, and jumping etc.

In the proposed model human activities are recognized like walking, running, jumping etc. The input dataset is analysed by extracting heir features and classified according to their specific characteristics.

The working of the proposed model (figure 3) described in the following steps:

Step 1: Input video dataset for human activity recognition.

Step 2: The video frame is then converted into frames or the images.

Step 3: Image buffer of initial frames are stored for the further processing.

Step 4: image frames are converted into space time shapes.

Step 5: the images are ready for pre-processing and feature extraction hence after pre-processing feature extraction is done.

Step 6: after preparing a model through deep learning CNN method the test dataset and compare the values for activity recognition and identify the activity

Step 7: finally Storing the resultant image frames for further usage.

The methodology of proposed model is comprising of four stages, i.e., converting the video frames into space time shapes and detecting the points, while pre-processing the image, then feature extraction and finally the classification

eISSN 1303-5150

(shown in Fig. 3)

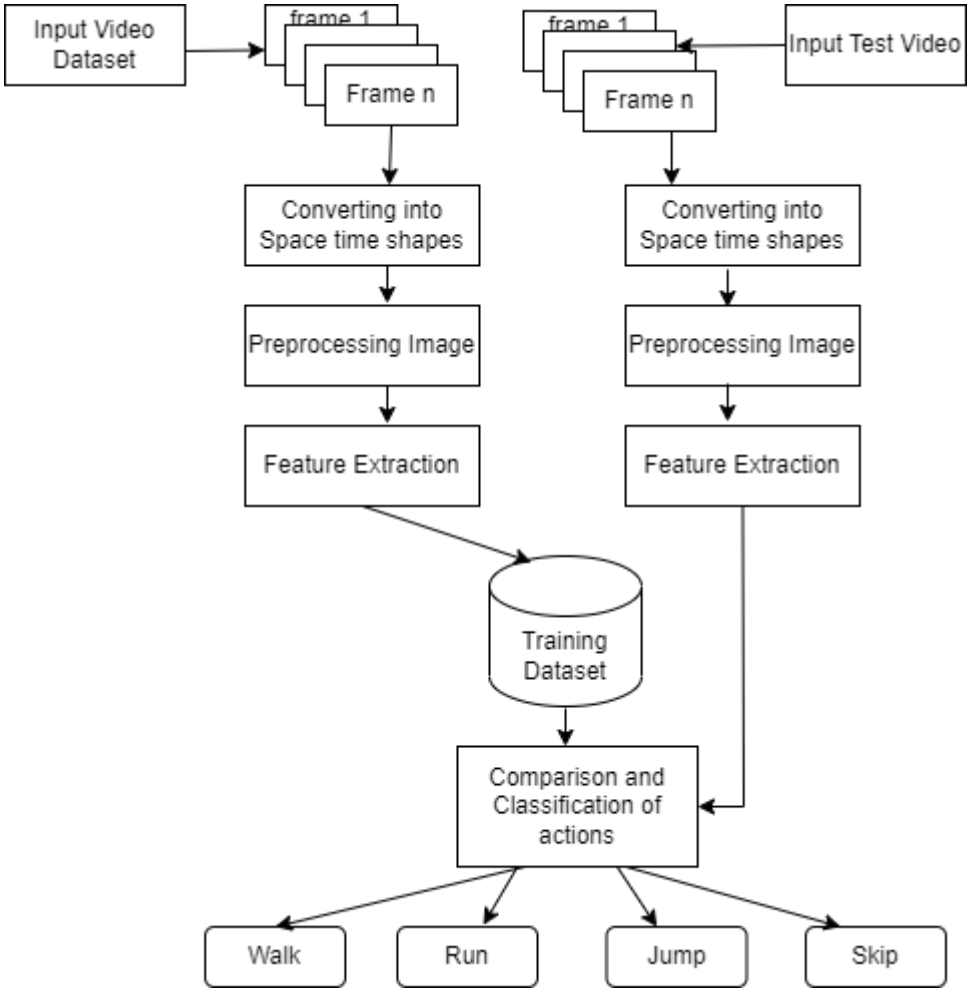

Figure 3: Architecture of proposed human activity recognition

5. Implementation& Results

The proposed Human activity recognition is implemented in Python. Data is handling using Pandas library. Scikit-learn library is used for the classification algorithms. Raw data can be classified by time series classification. This approach can be considered as end-to-end approach and because of it feature engineering is not required. Both the feature learning and classification is performed in one model. This approach helps to compare performance to the classification based on features. The model is developed in the Keras API in Python. Keras is a deep learning API which is running on the top of TensorFlow, which is a machine learning platform. In the model multi-layer perceptron is used which is a simple deep learning architecture and also used in the time series classification tasks(Abed & Rahman, 2017) [10].

Dataset Used

For the implementation of the proposed model publicly available dataset for action recognition is used. This dataset was initiated by KTH and it is one of the most standard datasets used for human action recognition. The KTH dataset contains six actions i.e., walking, running, jogging, boxing, hand waving and hand clapping. These actions

eISSN 1303-5150

are performed several times by 25 objects in different scenarios i.e., in outdoor and with the different scale or with different clothes also. The dataset contains 2391 sequences with homogeneous background with 25 fps frame rate. These sequences were down sampled to the resolution of 160 × 120 pixels with the average length of 4 seconds.

In the architecture of the model three hidden layers are present which are fully connected. The final layer for each class with one node is discriminative layer. Each node shows the probability of the input data which belongs to the class and the probability distribution is ensured by the previous layer.

Video sequences are provided to the algorithm which are performing different natural action like, run, jump and walk etc. Space time shapes of the different actions can be obtained by subtracting the background from each video sequence and using it in simple thresholding colour space.

In the proposed work, the CNN classifier is used for human action classification which leads to better results. The training-based classification consists of two different sets of actions' feature values that are extracted from the video frames.

For example, the walk actions' feature values are taken for training dataset and the testing video is taken for human action recognition. In the testing video, the features of the initial image frames are taken for classification.

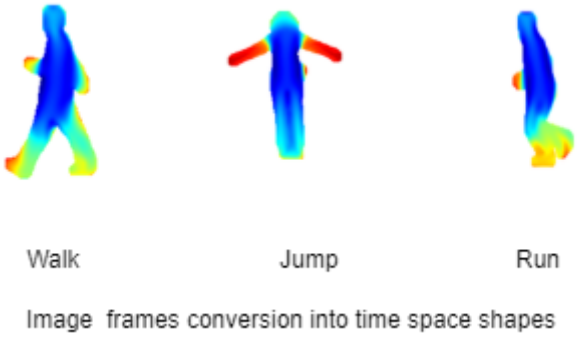

Figure 4: Time Space shapes of image frames

Table 1. classification for training (X) and testing datasets (Y)

X/Y	Walk	Run	Jump
Walk	Y	Y	N
Run	Y	Y	N
Jump	Y	Y	Y

The Table 1 Y Indicates correct classification of data, and N indicates wrong classification of data.

Table2 Confusion matrix using for training (X) and testing datasets (Y)

X/Y	Walk	Run	Jump
Walk	100	100	0
Run	68	100	0
Jump	100	100	100

eISSN 1303-5150

The result shows that the "walk" action has been classified correctly based on the training support vectors. But there are some confusions that arises in case of same feature values in the images. For example, run and jump actions have been classified wrongly, which means that some of the feature values are same in both actions, so that the confusion arises. As in the Table 2, the X and Y are two classes of inputs given for training-based classification. The first 20 frame features in the video are taken for training dataset. Based on this 20 frame feature values, the actions are classified. Y is the testing action dataset which is used for comparing the actions based on training video datasets.

6. Conclusion& Future Work

Human Action Recognition is one of the most popular and widely used method in computer vision and image processing. HAR is now used in diverse applications viz. surveillance, healthcare, managing public services etc. The advances in HAR technology are achievable through the development of Deep learning techniques. In the model proposed we used machine learning based classifiers based on convolutional neural network. The model can learn complex features directly from the raw data and recognize the activities. Experimental results reveal that the proposed deep learning model shows better performance compared to other approaches.

We have used Data Sources from KTH dataset, Weizmann dataset, for training and testing the human action recognition system. The given video is converted into sequences of jpeg images. A uniform size of 256x256 has been fixed as the standard size for the image frames taken for study. There are several Background Subtraction methods for identifying the foreground actions. In the research considers the static background video, for analysis. Deep learning(Gupta et al., 2021) [18] CNN classifier is used for classification of human actions(Lee & Lee, 2018) [19]. Traditional method, in some cases, fails to find the difference between the actions such as Jogging and Running. Hence, a better method (proposed) is needed for recognizing all kinds of actions in a perfect way. There are several methods available for extracting features from images. The features that are extracted play a vital role in recognizing human actions. Edge detectors, corner detectors and blob detectors are some of the methods used for extracting the features from the image. On testing all the methods, it has been observed that Harris Corner method (corner based) is best suitable for extracting the desired features from the images for the proposed study. In this research work, methods for recognizing human actions from a static background setting of controlled videos were explored. The proposed approach is used to analyse the human actions from the static view of background.

In the future work we will go further to study more efficient deep learning techniques to recognize more human actions. In the future the improved model for pattern detection from the weakly labelled dataset can be proposed with more evaluations in different situations for achieving better accuracy and efficiency.

References:

[01] Poppe, R. (2007). *Vision-based human motion analysis : An overview*. *108*, 4–18. https://doi.org/10.1016/j.cviu.2006.10.016

[02] Blank, M. (n.d.). *Actions as Space-Time Shapes*

eISSN 1303-5150

[03] Methods, A. R. (2019). *A Comprehensive Survey of Vision-Based Human Action Recognition Methods*. 1–20. https://doi.org/10.3390/s19051005

[04] Bodor, R. (n.d.). *Vision-Based Human Tracking and Activity Recognition.*

[05] Moeslund, T. B., Hilton, A., & Kru, V. (2006). *A survey of advances in vision-based human motion capture and analysis*. *104*, 90–126. https://doi.org/10.1016/j.cviu.2006.08.002

[06] Zhao, Z., & Zheng, P. (2012). *Object Detection with Deep Learning : A Review*. 1–21

[07] Ye, Q., Dong, J., & Zhang, Y. (2015). te d Ac ce p t. *Optik - International Journal for Light and Electron Optics*. https://doi.org/10.1016/j.ijleo.2015.08.103

[08] Pereira, F. C., Pereira, C. E., Pereira, F. C., Pereira, C. E., & Eduardo, C. (2015). *ScienceDirect Recognition of Cracks using UAVs Embedded Recognition of Cracks using UAVs for for Automatic Automatic Recognition Recognition of of Cracks Cracks using using UAVs UAVs*. 16–21. https://doi.org/10.1016/j.ifacol.2015.08.101

[09] Al-kaff, A., Mart, D., & Garc, F. (2017). *PT US CR*. https://doi.org/10.1016/j.eswa.2017.09.033

[10] Abed, A. A., & Rahman, S. A. (2017). *Python-based Raspberry Pi for Hand Gesture Recognition Python-based Raspberry Pi for Hand Gesture Recognition*. *September*. https://doi.org/10.5120/ijca2017915285

[11] Roobini, S., & Fenila Naomi, J. (2019). Smartphone Sensor Based Human Activity Recognition using Deep Learning Models. *International Journal of Recent Technology and Engineering*, *8*(1), 2740–2748

[12] Vahora, S. A., & Chauhan, N. C. (2019). Deep neural network model for group activity recognition using contextual relationship. *Engineering Science and Technology, an International Journal*, *22*(1), 47–54. https://doi.org/10.1016/j.jestch.2018.08.010

[13] Wang, K., He, J., & Zhang, L. (2019). Attention-based convolutional neural network for weakly labeled human activities' recognition with wearable sensors. *IEEE Sensors Journal*, *19*(17), 7598–7604. https://doi.org/10.1109/JSEN.2019.2917225

[14] Golestani, N., & Moghaddam, M. (2020). Human activity recognition using magnetic induction-based motion signals and deep recurrent neural networks. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-15086-2

[15] Zhou, X., Liang, W., Wang, K. I. K., Wang, H., Yang, L. T., & Jin, Q. (2020). Deep-Learning-Enhanced Human Activity Recognition for Internet of Healthcare Things. *IEEE Internet of Things Journal*, *7*(7), 6429–6438. https://doi.org/10.1109/JIOT.2020.2985082

[16] Mliki, H., Bouhlel, F., & Hammami, M. (2020). Human activity recognition from UAV-captured video sequences. *Pattern Recognition*, *100*, 107140. https://doi.org/10.1016/j.patcog.2019.107140

[17] Gupta, A., Anpalagan, A., Guan, L., & Khwaja, A. S. (2021). Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. *Array*, *10*(January), 100057. https://doi.org/10.1016/j.array.2021.100057

[18] Lee, D., & Lee, S. (2018). PT US CR. *Pattern Recognition*. https://doi.org/10.1016/j.patcog.2018.08.006