International Journal of Innovative Computing and Applications

ISSN online: 1751-6498 - ISSN print: 1751-648X
https://www.inderscience.com/ijica

HARDeep: design and evaluation of a deep ensemble model for human activity recognition

R. Raja` Subramanian, V. Vasudevan

DOI: 10.1504/IJICA.2023.10052593

Article History:

Received:	20 October 2021
Last revised:	02 March 2022
Accepted:	30 May 2022
Published online:	07 June 2023

HARDeep: design and evaluation of a deep ensemble model for human activity recognition

R. Raja Subramanian* and V. Vasudevan

Department of Computer Science and Engineering,
Kalasalingam Academy of Research and Education, India
Email: rajasubramanian.r@klu.ac.in
Email: vasudevan_klu@yahoo.co.in
*Corresponding author

Abstract: With the emergence of smartness in various fields including medical science, forensics and security, remote monitoring of human activities has gained more interests in research. The ambulatory health monitoring services includes monitoring the activities of mentally challenged and elderly people. In this research paper, we propose a novel framework for activity recognition from video sequences captured from static cameras and those captured from UAVs. The proposed framework, named HARDeep, consists of three models: an optional scene stabilisation model for UAV captured video sequences, a human detection model leveraging YOLOv3, and, to extract the set of video frames containing humans, an activity recognition model leveraging the ensemble of three deep learning models: GoogleNet, ResNet-50, and ResNet-101. HARDeep is evaluated against three datasets including Hollywood2, KTH and the UCF-ARG dataset, consisting of video sequences captured from UAVs. The recognition accuracies are compared with the various inference models leveraging wide learning paradigms.

Keywords: activity recognition; deep ensemble model; unmanned aerial vehicles; UAV; scene stabilisation; pretrained models; HARDeep; human detection; YOLO; fog computing.

Reference to this paper should be made as follows: Subramanian, R.R. and Vasudevan, V. (2023) 'HARDeep: design and evaluation of a deep ensemble model for human activity recognition', *Int. J. Innovative Computing and Applications*, Vol. 14, No. 3, pp.155–166.

Biographical notes: R. Raja Subramanian is the Assistant Professor of Department of Computer Science Engineering at Kalasalingam Academy of Research and Education, Tamil Nadu. He completed his Bachelor degree as Anna University Rank Holder in Computer Science and Master's degree as a Gold Medallist at Thiagarajar College of Engineering, Madurai. He has qualified in GATE-2016 and UGC-NET. He developed many consultancy and research projects for Esteemed Industries. He has published more than 30 research papers in journals and conferences. His areas of research include image/video processing, fog computing. He is a recipient of various awards from IUCEE for PBL and innovative teaching.

V. Vasudevan is a Registrar and Senior Professor of Kalasalingam Academy of Research and Education, Tamilnadu. He completed his PhD in Madurai Kamaraj University. His areas of interest include big data analytics, internet of things, and video processing. He has more than 34 years of teaching and research experience. He has more than 100 publications in national and international reputed journals. He received his DLitt award from International Economics. He was honoured with Dr. APJ Abdhul Kalam Award for Life Time Contribution in Teaching and Research. He received National Information Technology Awareness Award from Government of India. He supervised more than 25 PhD thesis.

1 Introduction

One of the prominent areas of research, finding application in ambulatory healthcare, surveillance of suspected people, remote monitoring of a particular place for unwanted activities, is human activity recognition (HAR). It includes identifying human movements, activities and behaviour from a remote area. Thus, assisting in possible analysis for doctors in treating mentally challenged people and for investigators in predicting intentions of criminals. With the development of unmanned aerial vehicles (UAV) over traditional CCTV surveillance cameras, the problems of object occlusion are reduced. Hence it is possible to remotely control the positions of the camera agents of UAV drone, overcoming the fixed area coverage of traditional cameras. HAR on video sequences captured by drones provides better utilities for monitoring in open spaces.

Table 1 Survey of various activity inference paradigms

Work	*Fog computing*	*Deep learning*	*Fixed camera inputs*	*UAV inputs*	*Performance analysis*			
					Latency	*Execution time*	*Jitter*	*Arbitration time*
Weighted hierarchical depth motion maps (Wang et al., 2016)		✓	✓					
Differential recurrent neural network (dRNN) (Veeriah et al., 2015)		✓	✓					
Layered 2-way RNN (Du et al., 2015)		✓	✓					
RNN + LSTM (Zhu et al., 2016)		✓	✓					
Convolutional nets (deep CNN) (Li et al., 2016)		✓	✓					
Object recognition through LieNet models (Huang et al., 2016)		✓	✓					
Convolutional nets (RNN + Deep CNN) (Shi et al., 2017)		✓	✓					
Pretrained Alex NET model with SVM classifier (AlDahoul et al., 2018)		✓	✓	✓				
Pretrained Alex NET model with softmax classifier (Mliki et al., 2019)		✓	✓	✓				
Two stream convolutional nets (Simonyan and Zisserman, 2014)		✓	✓					
LSTM + CNN (Ng et al., 2015)		✓	✓					
ResNET (Feichtenhofer et al., 2016)		✓	✓					
Volumetric video segmentation (Ke et al., 2005)	✓		✓					
Error correction output codes (ECOC) + SVM (Islam et al., 2019)	✓		✓					
Proposed work	✓	✓	✓	✓	✓	✓	✓	✓

Various state-of-the-art researches exist in the literature for recognising human activity from video sequences. But there is limited research in inferring the same from video sequences captured through drones. The video sequences captured through drones involve certain constraints including human presence in the scene, height of the camera, background lights, among others. In this paper, we depicted a novel approach, leveraging deep learning framework, for HAR, as an application to ambulatory healthcare service for mentally challenged patients. The problem is partitioned to two modules:

1 identification of useful frames in the video sequence

2 inference of appropriate human activity from the set of useful frames.

The useful frames correspond to those containing the human presence. The problem confined to drone-captured video sequences involves a scene stabilisation module as a preliminary step.

HAR typically finds services in crucial applications. In these applications, the results attained after the deadline, are not valuable, though being accurate. Processing the online video sequences obtained from drones typically require systems with sufficient storage and computation capabilities. Processing the videos in cloud poses additional data propagation time between edge devices to cloud and

vice versa. A fog-assisted cloud architecture overcomes this propagation delay, by providing fog nodes and servers close to the edge with minimum-required computation and storage powers. Processes that require more computation/ storage will be migrated to cloud. We propose fog-based architectures for human detection in video sequences and subsequently, activity inference from the human-present frames.

The subsequent sections of the paper are organised as follows: The state-of-the-art researches in HAR are discussed in Section 2. Section 3 depicts the proposed deep learning framework for activity inference. Experimental results are shown in Section 4, followed by conclusion and scope for future works.

2 Background and related works

Processing video sequences captured by drones, primarily require scene stabilisation. This pre-processing step is used to distinguish the movements caused by objects in the scene with the movements of the visual sensors in the drone. Scene stabilisation involves matching of features between consecutive frames followed by motion estimation and subsequent motion compensation. Feature matching involves detecting interest points or key points between two consecutive frames. Once matches are identified between frames, then motion detection is carried out. Motion can be global or local. Global motion corresponds to the motion induced by camera displacement, in case of drones. Whilst, the local motion is caused by object displacement in the video sequences. Motion compensation step is done for global motion detected sequences, to adjust the frames with respect to the previous position of the camera.

Researches to identify global motions (Hsiao et al., 2009; Shen et al., 2009) assumed that the foreground image is generally present in the centre block rather than the corners. This assumption is unrealistic in surveillance-based applications. Subsequent researches (Walha et al., 2015; Lowe, 2004) used scale invariant feature transforms (SIFT) to capture key points or interest points from the frames. Improvements on the SIFT leveraged algorithms are done using RANdom SAmple Consensus (RANSAC) (Fischler and Bolles, 1981) for filtering the outliers. Nearest neighbour criterion coupled with Euclidean distance measures is utilised to derive the displacement vectors. The vectors are analogous to the displacements occurred in the scene due to object passage or camera alignments. This is evident with the postulation that motion pertaining to object displacements is much faster than the motion created through camera displacements. Affine transformations are done on the frames to compensate the global motions. Other key point detectors including good features to track (GFTT) (Minaeian et al., 2018) are utilised to identify motions in consecutive frames. But these feature detectors have fixed parameters and are less reliable in detecting with global motions.

Various authors subjected the need to compute the magnitude and direction of the displacement variable in the frame sequences (Mliki et al., 2019). With this estimate, the global motion detection process becomes straightforward. Patterns recognised through optic flows provide such an estimate for displacement variables. With the implications of above surveys, researches (Burghouts et al., 2014) used the assumption that the presence of foreground images in the centre is more compared to that in the corners. In surveillance-based applications, model accuracy requirements render this assumption useless. The reliable interest point detector SIFT will be less efficient in time-critical applications (Walha et al., 2015). This is because SIFT and its extension, speeded-up robust features (SURF) are computationally complex. To overcome timing constraints, optical flow can be computed between two consecutive frames. Optical flow depicts the shape of the motion of objects in subsequent frames, owing to object or camera displacement. Analysing the shape of the motion distinguishes the variations caused by object and camera. Hence, the motion compensation step becomes unnecessary. Scene stabilisation in our proposed model uses optical flow computation leveraging Lucas-Kande algorithm, which is faster, simple and accurate (Barron et al., 1994; Khobragade et al., 2012).

The stabilised images are then subjected to the actual HAR framework. Images captured from fixed cameras can bypass the scene stabilisation process. In the next step, useful images need to be extracted from the set of video frames. The useful images are analogous to those with human presence, as human object is the area of interest in the given image, for the activity recognition task. The human detection step can be viewed as an object detection task. Several feature extraction techniques and deep learning models exist in the literature to perform object detection task. Binning techniques including histogram of oriented gradients (HOG) descriptors and SIFT are used for object detection (Burghouts et al., 2014; Uijlings et al., 2013). These techniques are robust against shape of the object under study. But these descriptors do not render faster results. Texture descriptors such as local binary patterns (LBP) and local ternary patterns (LTP), used in object detection are computationally expensive. These texture descriptors do not scale well with the number of images.

Various deep learning architectures are in exploration to perform human detection. Deep learning algorithms learn features automatically and provide effective and efficient results. Human detection algorithms leveraging convolution neural networks (CNNs) are prevalent in researches (He et al., 2016; Krizhevsky et al., 2012). The taxonomy of various object recognition, activity inference models is depicted in Table 1. Pretrained image net models (Wang et al., 2016; Veeriah et al., 2015; Du et al., 2015; Zhu et al., 2016; Li et al., 2016; Huang et al., 2016; Shi et al., 2017; Simonyan and Zisserman, 2014; Ng et al., 2015; Feichtenhofer et al., 2016; AlDahoul et al., 2018) including AlexNet, VGGNet, ResNet, LieNet, GoogleNet are leveraged for object inference paradigms by various researchers. Supervised CNN and Pretrained CNN models

are the two deep learning algorithms used by AlDahoul et al. (2018) on UCF-ARG dataset (Nagendran et al., 2010). It is empirically stated that pretrained models provide better results compared to the former models. Pretrained models are robust against typical visual parameters: camera altitude, lightings, angles, affecting video quality. These deep learning models are not stringent in close object scenarios. In our earlier research (Subramanian and Vasudevan, 2021), we have used deep genetic algorithms for recognising activities from human videos, streaming at a faster rate. Efficiency of the application is visualised leveraging fog computing frameworks. In this research, we strive to architect an ensemble deep learning network for HAR. Linear classifier, SVM (Islam et al., 2019), RNN (Veeriah et al., 2015; Du et al., 2015), and deep CNN (Wang et al., 2016; Zhu et al., 2016) models are the common classification methods used in HAR and object inference tasks. Typical deep learning models for object detection use pretrained model with the final layer replaced with a classifier, either SVM or softmax. Since, human detection is only the preliminary step and the main process is activity recognition, the time to extract the useful frames containing human, need to be much lower. Hence, we include a model for human object detection, that screens maximum frames in a limited duration compared to the above stated traditional object detection algorithms.

With the available set of useful frames, containing human images, the next step is to recognise their activity. Typical deep learning framework for activity recognition involves two phases: Extraction of spatio-temporal features from the frames, and inference of appropriate activity by analysing the extracted features. SURF leveraged algorithms are used by Islam et al. (2019) for HAR. The response times of activity recognition in the SURF based HAR are enhanced by blockchain-based fog computing architectures. The key point feature descriptors are less effective on image frames with noise. Hence a robust feature descriptor is required for detection and classification purposes. Wang et al. (2018) used CNN model to extract the spatial features from the frames. Subsequently time-domain features are extracted through long short-term memory models. Finally, the temporal features are merged with an optimisation layer, depicting the activity. Sargona et al. (2017) proposed a model leveraging the pretrained AlexNet model, with the last softmax layer replaced by the SVM-KNN hybrid classifier. The authors conducted evaluations on KTH and UCF-Sport datasets and proved the proficiency of pretrained models in recognising the human activity.

From the literature, it is evident that HAR can be done through UAV captured video sequences or fixed cameras in confined environment. Video sequences captured through drones require scene stabilisation process due to camera movements. The core part lies in extracting the useful frames containing human and inferring the activity of the user from consecutive frames. The key contributions of our work are as follows:

1 We propose a novel architecture for HAR from video sequences captured by static cameras or moving cameras.

2 An ensemble deep learning model for recognising human activity from the ordered set of video frames.

3 Evaluation of the algorithm against various benchmark datasets.

3 Proposed work

The proposed framework for HAR, HARDeep, consists of three modules:

1 an optional scene stabilisation module

2 identification of useful frames

3 inference of human activity.

The architecture of HARDeep is depicted in Figure 1.

3.1 Scene stabilisation

Scene stabilisation is carried out to distinguish the motions caused by object displacements and the motions caused by camera displacements. The latter motions require stabilisation to counter the motions caused by camera movements. If there exists a motion between two consecutive frames, an image pixel in initial frame (I_{K_t}) and that in the consecutive frame ($I_{K_{t+dt}}$), will have a displacement during the time interval *dt*. The pixel intensity during the motion is assumed to be constant and the images are therefore operated in greyscale. In order to estimate the displacement (*dx*, *dy*) of a pixel *I*(*x*, *y*, *t*) during the time interval *dt*, the Taylor series can be applied:

$$I(x+dx,\, y+dy,\, t+dt) = I(x,\, y,\, t) + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt + \cdots \quad (1)$$

Based on our assumption on the invariance of the pixel intensity between two consecutive frames, we can state that

$$I(x,\, y,\, t) = I(x+dx,\, y+dy,\, t+dt) \quad (2)$$

From (1) and (2)

$$\frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial x}dx + \cdots = 0 \quad (3)$$

The optical flow between the frames can be calculated by dividing (3) by *dt*.

$$-\frac{\partial I}{\partial t} = \frac{\partial I}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial I}{\partial y}\frac{\partial y}{\partial t} \quad (4)$$

where $V_x = \frac{\partial x}{\partial t}$ and $V_y = \frac{\partial y}{\partial t}$ are the optical flow components, with the direction $\vec{V}$ at the two-dimensional space. Hence the optic flow pattern can be calculated as:

$$-\frac{\partial I}{\partial t}=\frac{\partial I}{\partial x}V_x+\frac{\partial I}{\partial y}V_y \tag{5}$$

Rewriting (5) with respect to pixels m

$$-I_t(m)=I_x(m)V_x+I_y(m)V_y \tag{6}$$

With the known pixel intensities $I_x(m)$ and $I_y(m)$, the task is to solve (6) for V_x and V_y. Horn-Schunck and Lucas-Kande are the two commonly used techniques (Huang et al., 2016) to solve such optical flow equations. Lucas-Kande, being more efficient than the former in terms of time, is used in our experimentation. The techniques work on top of the assumption that in a window of dimension ($n \times n$), and centre m, all the pixels will have the same motion as pixel m. Hence, we obtain a system of n^2 equations, corresponding to various pixels p_i: $1 \le i \le n^2$ in the search window.

$$\begin{bmatrix} I_x(p_1) & I_y(p_1) \\ I_x(p_2) & I_y(p_2) \\ \vdots & \vdots \\ I_x(p_{n^2}) & I_y(p_{n^2}) \end{bmatrix}\begin{bmatrix} V_x \\ V_y \end{bmatrix}=-\begin{bmatrix} I_t(p_1) \\ I_t(p_2) \\ \vdots \\ I_t(p_{n^2}) \end{bmatrix} \tag{7}$$

There are n^2 equations to solve two unknowns in (7). Hence the equations can be solved using least square method.

$$\begin{bmatrix} V_x \\ V_y \end{bmatrix}=\begin{bmatrix} \sum_1^{n^2} I_x(p_i)^2 & \sum_1^{n^2} I_x(p_i)I_y(p_i) \\ \sum_1^{n^2} I_x(p_i)I_y(p_i) & \sum_1^{n^2} I_y(p_i)^2 \end{bmatrix}^{-1}\begin{bmatrix} -\sum_1^{n^2} I_x(p_i)I_t(p_i) \\ -\sum_1^{n^2} I_y(p_i)I_t(p_i) \end{bmatrix} \tag{8}$$

The motion vectors V_x and V_y of each pixel corresponding to potential motions regions and UAV displacements are estimated using (8). It is obvious that the velocity of displacement caused by object movements is greater than that caused by the camera displacements. The vectors are normalised with the displacement at highest magnitude. These norms are put together to form an image of appropriate pixel intensities. The resultant image will have pixels of intensity between [0..1]. Images with a smaller number of pixels are eliminated as noise. Hence the set of potential motion regions are detected.

3.2 Human detection model generation

The human detection task, being analogous to object detection in the video sequences, is carried out using you only see once (YOLOv3) network model (Lu et al., 2019; Zhang et al., 2019; Redmon and Farhadi, 2018). YOLOv3 leverages the idea of Redmon, in using residual learning of YOLOv2. YOLOv3 follows one-stage mode and multi-scale features. Thus, the model is faster and has high accuracy in object detection scenarios. It uses DarkNet53 unlike YOLOv2, which leverages on DarkNet19. DarkNet53 is a 53 layered convolution network. The series of 1 * 1 and 3 * 3 convolution layers compose the 53 layers of YOLOv3, including the fully connected layer, but leaving the residual layer. Each layer is followed by a batch normalisation (BN) layer and activated by the Leaky ReLU function. The presence of residual layers, derived from ResNet, covers the gradient explosion problems in the network and aids controlled gradient propagation and appropriate training. Despite having many convolution layers in the network, the number of parameters is reduced significantly. Thus, DarkNet53 serves in faster extraction of features from video frames.

Former versions of YOLO use softmax function in the final layer to classify the objects. But, YOLOv3 reforms the network using independent logistic classifiers, in place of a softmax classifier. Thus, the probability of the object to map any one of the class labels is calculated. The replacement of softmax layer, allows usage of cross-entropy for classification loss calculation. Thus, decreasing the computation complexity of the network. YOLOv3 is known for its fast detection. HAR, being a time-critical application, preliminarily requires effective and efficient recognition of video frames with human objects. YOLOv3, the state-of-the-art object detector can perform this task accurately and at a faster rate.

YOLO models typically scan the images on the whole and partitions the images into various windows of size s * s. For each object (human object in our scenario) and corresponding windows, the model calculates the probability that the window composes the centre of the object. The object is categorised with appropriate confidence, if the probability crosses certain threshold. Windows are confined to B boundary boxes and the confidence levels are estimated for each box simultaneously. Confidence level depicts the inference of object in the corresponding bounding box. The probability of this confidence parameter P_C is calculated as follows:

$$P_C = P(object) * IOU \tag{9}$$

$$IOU=\frac{area(BB_r \cap BB_d)}{area(BB_r \cup BB_d)} \tag{10}$$

$P(object)$ is the probability that the bounding box contains the object. Intersection over union (IOU) depicts the accuracy of an object detector on the dataset in question. BB_r represents the reference bounding box corresponding to training labels. BB_d is the decision bounding box.

Each bounding box compose a configuration (x, y), h, w, P_C where (x, y) is the centre pixel, h, w depicts the height, width respectively and P_C represents the confidence level. For each grid, the probability of an object in the frame to get classified in one of the classes C is calculated as $P(C_i|object)$. With respect to the confidence level, the frames which compose the object corresponding to class: person are categorised as useful frames containing human. As the bounding box which consists of maximum probability of containing a class object is indicated as the object holder, the complexity of the model is not affected by the overlapping objects.

Figure 1 HARDeep architecture

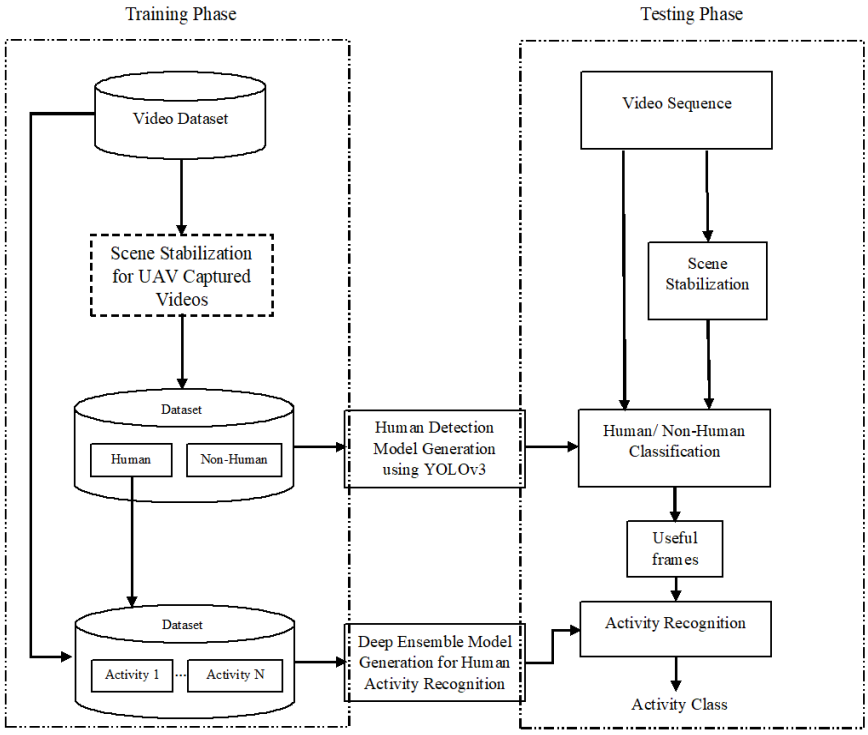
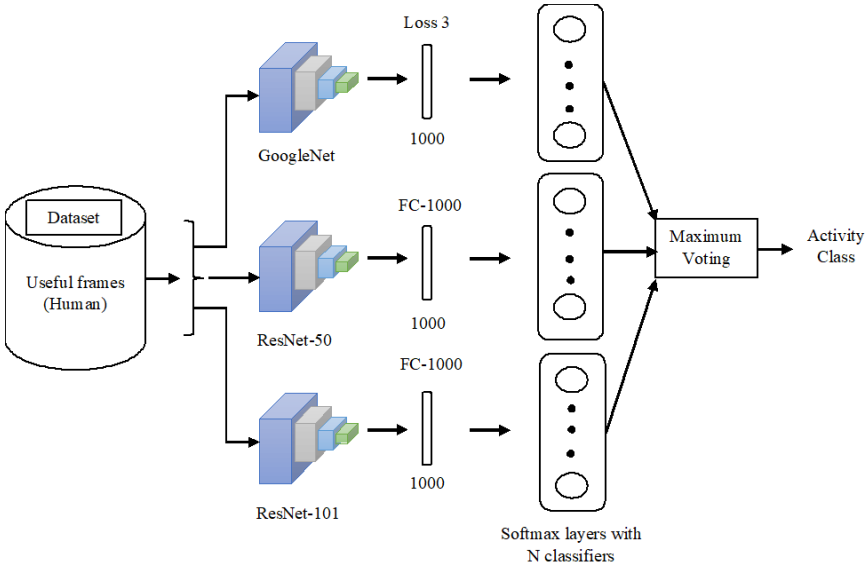


Figure 2 Ensemble deep model for HAR (see online version for colours)



3.3 Deep ensemble model for activity recognition

An ensemble model combines the efforts of multiple learning models and accumulates the results by weighting their decisions. Thus, ensemble algorithm will have better performance than the atomic ones. In order to build a deep ensemble model, the performance of five pretrained models: Alexnet, Vgg-16, Resnet-50, Resnet-101 and GoogleNet are evaluated and the best three models: Resnet-50, Resnet-101, GoogleNet are collaborated to develop an ensemble classifier leveraging majority voting scheme as shown in Figure 2. These ImageNet models have a common capability of describing an image with 1000 features in one of the layers. The layer from which the features are extracted are as follows: Loss3 in GoogleNet, FC-1000 in Resnet-50 and Resnet-101. The ImageNet models being pretrained and successful in classifying various objects, are subjected to high variance issues. The issue propagates with the increase in the number of trainings. The input for the activity recognition model in HARDeep comes from the output of YOLO human detection model. The human frames bounded at YOLO come in huge numbers as input to activity recognition model. Hence the model has a higher possibility to get overfitted. To overcome the variance issues, ensemble of three best ImageNet models is leveraged.

GoogleNet composes a complex structure encapsulating various convolutional and pooling layers, and block stacks. The image features are stacked across various networks. In order to propagate the errors in the deep network, there softmax layers are used during training. Resnet consists of a deep network architecture with a novel residual blocks that project input with appropriate identity mapping function to result in the appropriate output criterion. Resnet architectures are highly significant in overcoming the vanishing gradient issues, which lasts due to the high variance of the model. Hence HARDeep – activity recognition model ensembles two Resnet architectures vis. Resnet-50 and Resnet-101, along with the GoogleNet architecture. Hence the model issues with respect high variance, overfitting and vanishing gradient are overcome. The final layers of the three pretrained deep models are replaced with a softmax layer with N classes, where N corresponds to the activity classes. The ensemble classifier outputs a class that receives a maximum vote from the deep models. The decision of the k^{th} deep classifier is depicted as $dcs(c, a) \in (0, 1)$, $c = \{1, 2, 3, \ldots, P\}$ and $a = \{1, 2, 3, \ldots, N\}$ where P corresponds to the number of classifiers ($P = 3$ in the proposed model) and N, the number of activity classes. If the k^{th} classifier outputs the class a_i, then $dcs(c, a_i) = 1$ and 0, otherwise; $i \in k$. The majority voting criterion is put forth as follows (11):

$$\sum_{c=1}^{P} dcs(c, a) = \max_{a=1}^{N} \sum_{c=1}^{N} dcs(c, a_i) \tag{11}$$

Such an ensemble model outperforms atomic models in terms of response time. But the operational complexity of the model is higher. To overcome this, HARDeep works on top of a fog computing architecture fogbus. The orchestration of HARDeep in fogbus (Tuli et al., 2019) and the empirical results on the performance is depicted in Section 5.

4 Experimental study

The experimentation is carried out in two phases: human detection from video sequences and activity recognition from the detected set of useful frames. Three benchmark datasets used, are described and the series of experiments conducted on those datasets are explained in this section. The empirical evaluations of the proposed approach are compared with the state-of-the-art approaches.

4.1 Hollywood2 action dataset

Hollywood2 action dataset (Marszalek et al., 2009) is a dataset captured from fixed camera. Thus, these video sequences do not require scene stabilisation, to counter the camera movements. The dataset composes 1,707 video sequences from around 69 movies. The video sequences evidence 12 different activity classes including answer_phone, handshake, drive_car, eat, get_out_of_car, fight_person, hug_person, kiss, run, sit_down, sit_up, and stand_up. The video dataset occupies storage of around 15 GB playing for 20.1 hours. Figure 3 depicts the image frames from Hollywood2 dataset pertaining to various activities.

Figure 3 Sample frames depicting actions: Hollywood2 dataset (see online version for colours)

4.2 KTH dataset

KTH dataset (Schuldt et al., 2014) also composes video sequences captured from static cameras. This dataset, commonly used for activity recognition research, has 2391 video sequences captured from 25 fps camera in same background. The video is contributed by 25 people with six activities namely boxing, clapping, hand-waving, jogging, running, walking. The sample images from KTH dataset for the six classes of activities are depicted in Figure 4.

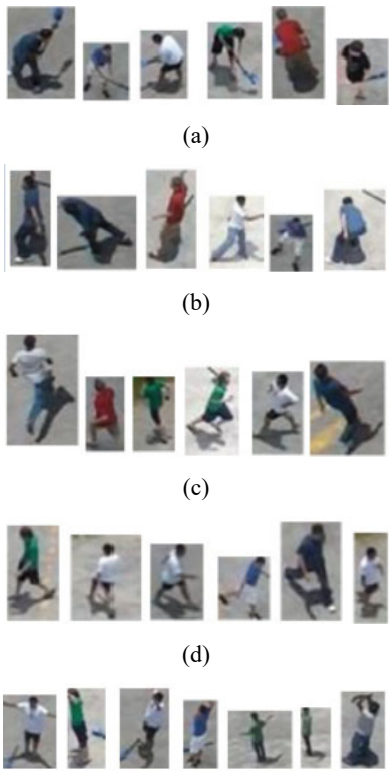
Figure 4 Sample frames depicting actions: KTH dataset (see online version for colours)

4.3 UCF-ARG dataset

UCF-ARG dataset (Nagendran et al., 2010) is a benchmark dataset for conducting experiments on human detection and corresponding activity recognition from video sequences captured from aerial cameras. The dataset possesses videos contributed by 12 people witnessing ten human activities. The dataset is multifaceted with its constraints including the extreme ego-motion, the lighting levels and the camera

height variation. In our experimentation, we focus on the recognition of following five activity classes: throwing, running, digging, waving and walking. The sample image frames from the UCF-ARG dataset for the five activity classes are depicted in Figure 5.

Figure 5 Sample frames depicting actions: UCF-ARG dataset (a) digging, (b) throwing, (c) running, (d) walking and (e) waving (see online version for colours)



(a)

(b)

(c)

(d)

(e)

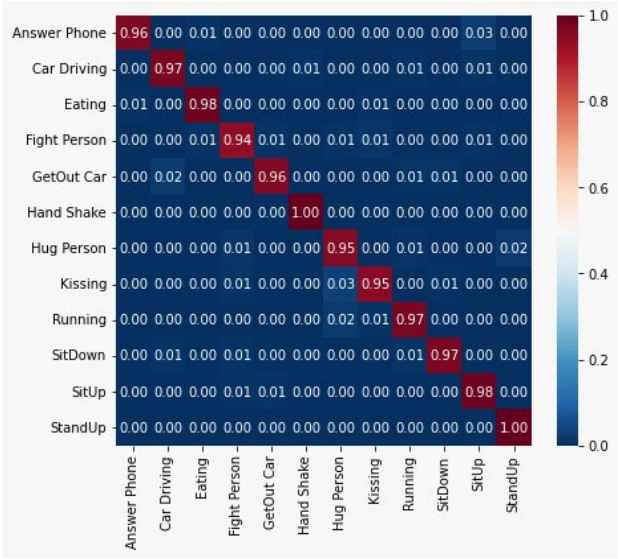
4.4 Human detection from video sequences

Video sequences captured by dynamic cameras in drones are subjected to scene stabilisation. The stabilised video sequences are then subjected to the first series of experiments to extract the useful frames. The useful frames are analogous to those set of frames containing human. The proposed framework uses YOLOv3 model to detect human objects in the video sequences. The video frames are parted into various windows, and the probability for the presence of the centre of human object in the window is calculated. The appropriate number of objects is extracted through the calibrated B boundary boxes. The YOLOv3 based human detection model showed a detection accuracy of 99.87% on Hollywood2, 99.94% on KTH and 99.72% UCF-ARG datasets respectively. It is obvious that the human detection using YOLOv3 outperforms the techniques leveraging the CNN models with the final layer replaced with SVM or softmax layer. In addition, YOLO model is much faster than the CNN models for object detection tasks.

4.5 HAR from useful frames

The detected set of useful frames containing human is used for subsequent recognition of activities. All the deep ensemble models are configurated with 70% Training and 30%. The video dataset consisting the activities and the human frame are subjected to the ensemble models composing GoogleNet, ResNet-50 and ResNet-101. The softmax layer, forming the last layer of the ImageNet models will consist of the number of activities to be recognised (N). Furthermore, we have a majority voting criterion to take the best of the three ImageNet models. Hence this ensemble model exhibits better recognition rates compared to the state-of-the-art HAR techniques. In order to have a trade of between the accuracy and complexity of ensemble model, we have experimented the activity recognition framework leveraging fog computing. The captured videos from UAV or static cameras are stored in the fog master node. The human detection model present in the fog master generates the set of useful human frames. These human frames are provided as input to three cluster heads, each containing GoogleNet, ResNet-50 and ResNet-101 models with the appropriate softmax layers. The activity recognition tasks are carried out in parallel by the three nodes consisting of the ImageNet models. The results from the fog nodes are ensemble at the fog master with the majority voting criterion.

Figure 6 HARDeep in Hollywood2 dataset: confusion matrix (see online version for colours)



The inference accuracy of the HARDeep framework is evaluated against various standard and complex video datasets including Hollywood2 dataset, KTH dataset and the UCF-ARG dataset, containing videos captured from UAV. Table 2 depicts the evaluation results of HARDeep on the three activity datasets under study. The confusion matrix for Hollywood2 dataset is depicted in Figure 6. It exhibits a recognition accuracy of 97.13%. The proposed framework recognises activities from KTH dataset at the accuracy level of 96.75%. The confusion matrix for the KTH dataset is presented in Figure 7. The framework is also experimented

against dynamic aerial camera captured video sequences from UCF-ARG dataset, and an inference accuracy of 80.72% is obtained. Figure 8 consists of the confusion matrix for the UCF-ARG dataset. The proficiency of HARDeep is compared with various activity inference techniques and the same is depicted in Table 3. It shows that the proposed model exhibits better recognition accuracy than the existing models. These results are mainly because of the combination of three best image classifiers and the YOLOv3 model, effectively providing the human-frames as input to these models. The accuracy of HARDeep is improved greatly with respect to Hollywood2 and UCF-ARG datasets, compared to the state-of-the-art models. The improvisation is witnessed with the sophistication of HARDeep architecture with ensemble deep learning models. However, the ensemble models, does not have significant effect in datasets like KTH, compared to state-of-the-art models. KTH consists of images captured in homogeneous backgrounds and variations in light. The most similar among the six actions like jog, run, walk are also distinguished through the geometrical difference in leg part of the image. Hence the atomic models in the HARDeep architecture typically give identical results, thus making the majority voting straight forward.

Figure 7 HARDeep in KTH dataset: confusion matrix (see online version for colours)

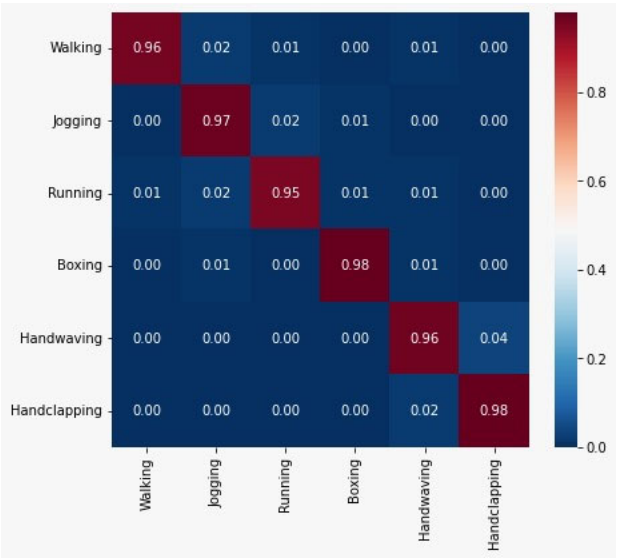


Table 2 Evaluation results of HARDeep on three activity datasets under study

Datasets	*TP*	*TN*	*FN*	*FP*
Hollywood2	3,282	83,588	1,406	1,179
KTH	18,939	79,198	1,313	1,995
UCF-ARG	14,895	57,033	9,035	8,087
Datasets	*Precision*	*Recall*	*Accuracy*	
Hollywood2	0.7357	0.7001	0.9713	
KTH	0.9047	0.9352	0.9675	
UCF-ARG	0.6481	0.6224	0.8072	

Figure 8 HARDeep in UCF-ARG dataset: confusion matrix (see online version for colours)

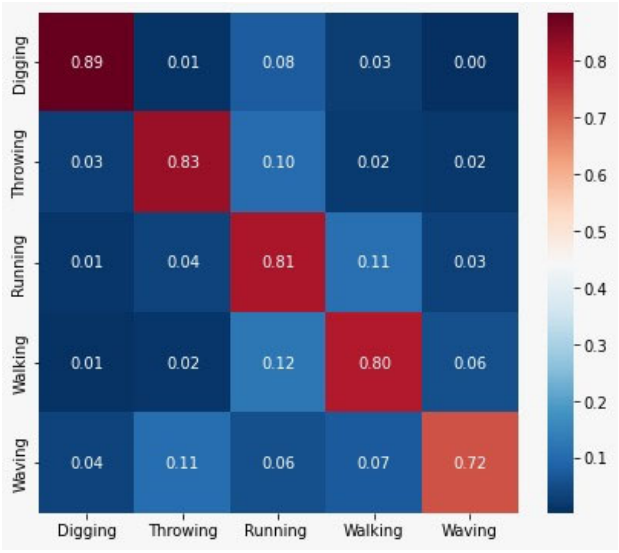


Table 3 Comparison of HARDeep with various HAR models

Techniques	*Hollywood2*	*KTH*	*UCF-ARG*
Alex NET model with softmax classifier (Mliki et al., 2019)	-	-	68.00
HoG + SIFT. (Burghouts et al., 2014)	-	-	57.00
ECOC based multi-class SVM (Islam et al., 2019)	87.00	75.00	-
Multi-skip feature stacking (Lan et al., 2015)	68.00	-	-
HoG + Gaussian classifier (Tian et al., 2012)	-	94.50	-
Spatio-temporal relationship match (Ryoo and Aggarwal, 2009)	-	91.10	-
Interest points detection + clustering (Bregonzio et al., 2009)	-	93.17	-
Sparse Bayesian feature classifier (Thi et al., 2012)	-	94.67	-
Bag of video words (Roshtkhari and Levine, 2013)	-	95.00	-
Local features + randomised KD trees (Mikolajczyk and Uemura, 2011)	-	95.3	-
Volumetric features (Ke et al., 2005)	64.60	-	-
Proposed model	97.13	96.75	80.72

5 Performance evaluation of the HARDeep model in fog environment

The proposed model, HARDeep, for activity recognition gives better accuracy with three standard video datasets, vis.

Hollywood2, KTH and UCF-ARG. Since activity recognition models finds use cases in time-critical applications, the model complexity in terms of response time need to be estimated. We execute the ensemble HAR model in fogbus (Lan et al., 2015; Tian et al., 2012) framework. Fogbus, being the state-of-the-art framework for modelling fog-based cloud computing architectures, is leveraged. The efficiency of the proposed HAR model is estimated in terms of latency, execution time, jitter and arbitration rate, for four different fog scenarios:

1 master only

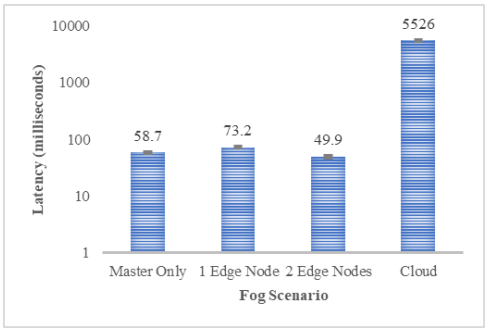
2 single edge node

3 two edge nodes

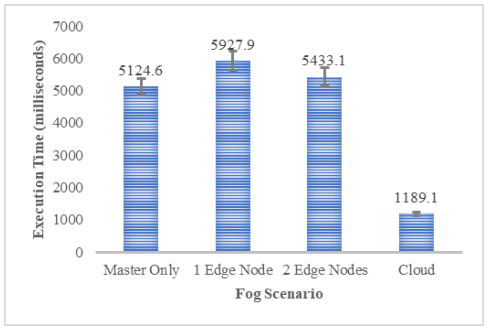
4 cloud.

Fog assisted models performs at its best, when the requests from the app is handled at the fog nodes, rather being forwarded to cloud. Fog shows its effectiveness by diminishing the unwanted data transit time to the cloud. It is evident from Figure 9(a) that latency is higher for the requests handled at the cloud layer, due to the multi-hop data transfer requirements. Whilst, the latency for the task handled at master or with any of the edge nodes is lower than that of the cloud. This is due to the fact that all the communications are confined through single hop.

Figure 9 (a) Latency and (b) execution time in different fog setups (see online version for colours)



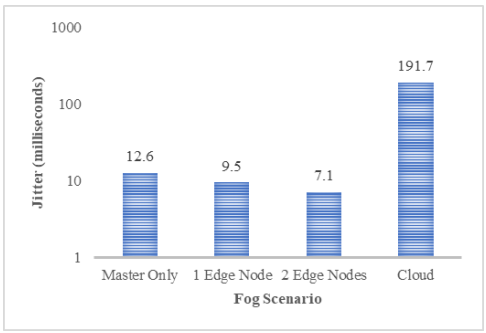
(a)



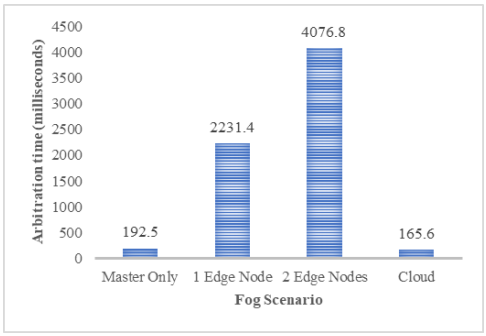
(b)

On the other hand, the sophisticated processing capabilities of the cloud render the task to get executed at a much faster rate. The fog worker nodes composing processing capacity with low clock frequency, exhibits a higher execution time. Figure 9(b) depicts the execution characteristics of the HARDeep model under four different fog scenarios. The delay in response for consequent job requests is measured by jitter. The jitter characteristics of the proposed HARDeep model under various fog setups are presented in Figure 10(a). It is obvious that the response time for the tasks is affected by the data transit time for the data being processed at cloud. For local nodes, the jitter is low and the measure is directly proportional to the number of edge nodes contained in the cluster. Load balancing, task assignment, parallelisation becomes difficult to manage with the increase in number of nodes for consensus. Hence arbitration rate is very low at the master only and the cloud scenarios. And it increases with the increase in number of nodes in the network. The arbitration characteristics of the HARDeep model at different fog setups are presented in Figure 10(b).

Figure 10 (a) Jitter and (b) arbitration time in different fog setups (see online version for colours)



(a)



(b)

6 Conclusions

Activity recognition is evolving as a strong area of research in the area of medical science, forensics and security. Giving more focus to medical science, the recognition of the daily activities of mentally challenged and elderly people, from a remote location becomes predominant. In the view of such time-critical task, we have proposed models for inference of human objects and subsequent identification of human activity from the videos captured from the static cameras and those from the dynamic UAVs. The human detection model is constructed on top of YOLOv3 object detector and a fog assisted deep ensemble classifier is leveraged for activity recognition tasks. The model is

evaluated against standard video datasets and a sound recognition accuracy is evidenced. In future, we aspire to develop a model to recognise the exercises, suggested by physiotherapists for patients with musculoskeletal problems. Thus, assisting the physicians to monitor the patient from remote locations and provide appropriate suggestions online.

References

AlDahoul, N., Sabri, M., Qalid, A. and Mansoor, A.M. (2018) 'Real-time human detection for aerial captured video sequences via deep models', *Computational Intelligence and Neuroscience*, Vol. 2018, No. 1, pp.1–14.

Barron, J.L., Fleet, D.J. and Beauchemin, S.S. (1994) 'Performance of optical flow techniques', *International Journal of Computer Vision*, Vol. 12, No, 1, pp.43–77.

Bregonzio, M., Shaogang, G. and Tao, X. (2009) 'Recognising action as clouds of space-time interest points, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.1948–1955.

Burghouts, G., van Eekeren, A. and Dijk, J. (2014) 'Focus-of-attention for human activity recognition from uavs', in *Electro-Optical and Infrared Systems: Technology and Applications XI*, International Society for Optics and Photonics, Vol. 9249, p.92490T.

Du, Y., Wang, W. and Wang, L. (2015) 'Hierarchical recurrent neural network for skeleton based action recognition', in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1110–1118, Boston, MA, USA.

Feichtenhofer, C., Pinz, A. and Wildes, R.P. (2016) *Spatiotemporal Residual Networks for Video Action Recognition*, CoRR, abs/1611.02155.

Fischler, M.A. and Bolles, R.C. (1981) 'Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography', *Communications of the ACM*, Vol. 24, No. 6, pp.381–395.

He, K., Zhang, X., Ren, S. and Sun, J. (2016) 'Deep residual learning for image recognition', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.770–778.

Hsiao, J.P., Hsu, C-C., Shih, T.C., Hsu, P.L., Yeh, S.S. and Wang, B.C. (2009) 'The real-time video stabilization for the rescue robot', in *2009 ICCAS-SICE*, IEEE, pp.4364–4369.

Huang, Z., Wan, C., Probst, T. and Gool, L.V. (2016) *Deep Learning on Lie Groups for Skeleton-Based Action Recognition*, arXiv Prepr, Cornell University Library, Ithaca, NY, USA.

Islam, N., Faheem, Y., Ud Din, I., Talha, M., Guizani, M. and Khalil, M. (2019) 'A blockchain-based fog computing framework for activity recognition as an application to e-Healthcare services', *Future Generation Computer Systems*, Vol. 100, No. 1, pp.569–578.

Ke, Y., Sukthankar, R. and Hebert, M. (2005) 'Efficient visual event detection using volumetric features', in *Tenth IEEE International Conference on Computer Vision, ICCV'05*, Vol. 1, pp.166–173.

Ke, Y., Sukthankar, R. and Hebert, M. (2005) 'Efficient visual event detection using volumetric features', in *Tenth IEEE International Conference on Computer Vision, ICCV'05*, Vol. 1, pp.166–173.

Khobragade, A., Kulat, K. and Dethe, C. (2012) 'Motion analysis in video using optical flow techniques', *International Journal of Information Technology and Knowledge Management*, Vol. 5, No. 1, pp.9–12.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) 'Imagenet classification with deep convolutional neural networks', in *Advances in Neural Information Processing Systems*, pp.1097–1105.

Lan, Z., Lin, M., Li, X., Hauptmann, A.G. and Raj, B. (2015) 'Beyond Gaussian pyramid: multi-skip feature stacking for action recognition', in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp.204–212.

Li, Y., Li, W., Mahadevan, V. and Vasconcelos, N. (2016) 'Vlad3: encoding dynamics of deep features for action recognition', in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1951–1960, Las Vegas, NV, USA.

Lowe, D.G. (2004) 'Distinctive image features from scale-invariant keypoints', *International Journal of Computer Vision*, Vol. 60, No. 2, pp.91–110.

Lu, S., Wang, B., Wang, H., Chen, L., Linjian, M. and Zhang, X. (2019) 'A real-time object detection algorithm for video', *Computers and Electrical Engineering*, Vol. 77, No. 1, pp.398–408.

Marszalek, M., Laptev, I. and Schmid, C. (2009) 'Actions in context', in *IEEE Conference on Computer Vision and Pattern Recognition*.

Mikolajczyk, K. and Uemura, H. (2011) 'Action recognition with appearance-motion features and fast search trees', *Comput. Vision Image Underst.*, Vol. 115, No. 3, pp.426–438.

Minaeian, S., Liu, J. and Son, Y.J. (2018) 'Effective and efficient detection of moving targets from a uavs camera', *IEEE Transactions on Intelligent Transportation Systems*, Vol. 19, No. 2, pp.497–506.

Mliki, H., Bouhlel, F. and Hammami, M. (2019) 'Human activity recognition from UAV-captured video sequences', *Pattern Recognition*, Vol. 100, No. 1, pp.107–140.

Nagendran, A., Harper, D. and Shah, M. (2010) 'UCF-ARG dataset', University of Central Florida [online] http://crcv.ucf.edu/data/UCF-ARG.php (accessed 8 November 2022).

Ng, J.Y-H., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R. and Toderici, G. (2015) 'Beyond short snippets: deep networks for video classification', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp.4694–4702.

Redmon, J. and Farhadi, A. (2018) *Yolov3: An Incremental Improvement*, arXiv preprint arXiv: 1804.0276.

Roshtkhari, M.J. and Levine, M.D. (2013) 'Human activity recognition in videos using a single example', *Image and Vision Computing*, Vol. 31, No. 1, pp.864–876.

Ryoo, M.S. and Aggarwal, J.K. (2009) 'Spatio-temporal relationship match: video structure comparison for recognition of complex human activities', *IEEE International Conference on Computer Vision (ICCV)*, pp.1593–1600.

Sargano, A.B., Wang, X., Angelov, P. and Habib, Z. (2017) 'Human action recognition using transfer learning with deep representations', in *2017 International Joint Conference on Neural Networks (IJCNN)*, IEEE, pp.463–469.

Schuldt, C., Laptev, I. and Caputo, B. (2014) 'Recognizing human actions: a local SVM approach', in *Proceedings of the 17th International Conference on Pattern Recognition, ICPR 2004*, Vol. 3, pp.32–36.

Shen, H., Pan, Q., Cheng, Y. and Yu, Y. (2009) 'Fast video stabilization algorithm for UAV', in *2009 IEEE International Conference on Intelligent Computing and Intelligent Systems*, IEEE, Vol. 4, pp.542–546.

Shi, Y., Tian, Y., Wang, Y. and Huang, T. (2017) 'Sequential deep trajectory descriptor for action recognition with three-stream CNN', *IEEE Transactions on Multimedia*, Vol. 19, No. 7, pp.1510–1520.

Simonyan, K. and Zisserman, A. (2014) 'Two-stream convolutional networks for action recognition in videos', in *Advances in Neural Information Processing Systems*.

Subramanian, R.R. and Vasudevan, V. (2021) 'A deep genetic algorithm for human activity recognition leveraging fog computing frameworks', *Journal of Visual Communication and Image Representation*, Vol. 77, No. 1, pp.103–132.

Thi, T.H., Cheng, L., Zhang, J., Wang, L. and Satoh, S. (2012) 'Integrating local action elements for action analysis', *Comput. Vision Image Underst.*, Vol. 116, No. 3, pp.378–395.

Tian, Y., Cao, L., Liu, Z. and Zhang, Z. (2012) 'Hierarchical filtered motion for action recognition in crowded videos', *IEEE Trans. Syst. Man Cybern.*, Vol. 42, No. 3, pp.313–323.

Tuli, S., Redowan M., Shikhar T. and Rajkumar B. (2019) 'FogBus: a blockchain-based lightweight framework for edge and fog computing', *Journal of Systems and Software*, Vol. 154, No. 1, pp.22–36.

Uijlings, J.R., Van De Sande, K.E., Gevers, T. and Smeulders, A.W. (2013) 'Selective search for object recognition', *International Journal of Computer Vision*, Vol. 104, No. 2, pp.154–171.

Veeriah, V., Zhuang, N. and Qi, G.J. (2015) 'Differential recurrent neural networks for action recognition', in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp.4041–4049, Santiago, Chile.

Walha, A., Wali, A. and Alimi, A.M. (2015) 'Video stabilization with moving object detecting and tracking for aerial video surveillance', *Multimedia Tools and Applications*, Vol. 74, No. 17, pp.6745–6767.

Wang, L., Xu, Y., Cheng, J., Xia, H., Yin, J. and Wu, J. (2018) 'Human action recognition by learning spatiotemporal features with deep neural networks', *IEEE Access*, Vol. 6, No. 1, pp.17913–17922.

Wang, P., Li, W., Gao, Z., Zhang, J., Tang, C. and Ogunbona, P.O. (2016) 'Action recognition from depth maps using deep convolutional neural networks', *IEEE Transactions on Human-Machine Systems*, Vol. 46, No. 4, pp.498–509.

Zhang, Y., Shen, Y. and Zhang, J. (2019) 'An improved tiny-yolov3 pedestrian detection algorithm', *Optik*, Vol. 183, No. 1, pp.17–23.

Zhu, W., Lan, C., Xing, J. et al. (2016) *Co-Occurrence Feature Learning for Skeleton Based Action Recognition Using Regularized Deep LSTM Networks*, arXiv Prepr, AAAI, Phoenix, Arizona, USA, Vol. 2, p.8.