IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 29, 2022, accepted June 8, 2022, date of publication June 13, 2022, date of current version June 20, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3182315

A Comparison Between Various Human Detectors and CNN-Based Feature Extractors for Human Activity Recognition via Aerial Captured Video Sequences

NOUAR ALDAHOUL[1,2], HEZERUL ABDUL KARIM[2], (Senior Member, IEEE), AZNUL QALID MD. SABRI[1], (Senior Member, IEEE), MYLES JOSHUA TOLEDO TAN[3], (Member, IEEE), MHD. ADEL MOMO[2], AND JAMIE LEDESMA FERMIN[3], (Student Member, IEEE)

[1]Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur 50603, Malaysia
[2]Faculty of Engineering, Multimedia University, Cyberjaya 63100, Malaysia
[3]College of Engineering and Technology, University of St. La Salle, Bacolod 6100, Philippines

Corresponding author: Nouar Aldahoul (nouar.aldahoul@live.iium.edu.my)

This work was supported in part by Multimedia University, and in part by the University Malaya Research Grant (Faculty Program) under Grant GPF095A-2020.

ABSTRACT Human detection and activity recognition (HDAR) in videos plays an important role in various real-life applications. Recently, object detection methods have been used to detect humans in videos for subsequent decision-making applications. This paper aims to address the problem of human detection in aerial captured video sequences using a moving camera attached to an aerial platform with dynamical events such as varied altitudes, illumination changes, camera jitter, and variations in viewpoints, object sizes and colors. Unlike traditional datasets that have frames captured by a static ground camera with medium or large regions of humans in these frames, the UCF-ARG aerial dataset is more challenging because it contains videos with large distances between the humans in the frames and the camera. The performance of human detection methods that have been described in the literature are often degraded when input video frames are distorted by noise, blur, illumination changes, and the like. To address these limitations, the object detection methods used in this study were trained on the COCO dataset and evaluated on the publicly available UCF-ARG dataset. The comparison between these detectors was done in terms of detection accuracy. The performance evaluation considers five human actions (digging, waving, throwing, walking, and running). Experimental results demonstrated that EfficientDetD7 was able to outperform other detectors with 92.9% average accuracy in detecting all activities and various conditions including blurring, addition of Gaussian noise, lightening, and darkening. Additionally, deep pre-trained convolutional neural networks (CNNs) such as ResNet and EfficientNet were used to extract highly informative features from the detected and cropped human patches. The extracted spatial features were utilized by Long Short-Term Memory (LSTM) to consider temporal relations between features for human activity recognition (HAR). Experimental results found that the EfficientNetB7-LSTM was able to outperform existing HAR methods in terms of average accuracy (80%), and average F1 score (80%). The outcome is a robust HAR system which combines EfficientDetD7, EfficientNetB7, and LSTM for human detection and activity classification.

INDEX TERMS Aerial captured video, convolutional neural network, human activity recognition, human detection, long short-term memory, transfer learning.

The associate editor coordinating the review of this manuscript and approving it for publication was Juan Liu.

I. INTRODUCTION

Human detection is a computer task that has been in development for at least two decades now [1]. This technology

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

has had various applications in search and rescue [2]–[8], law enforcement [9]–[11], pedestrian detection for traffic management and automated driver assistance [12]–[19], fall detection [20]–[25], and many other functions, including the decision-making steps that ensue [26]. And oftentimes, the technology is deployed through unmanned aerial vehicles (UAVs) [4], [6]–[8], [27]–[29] given their flexibility, longer range of tracking, and ability to acquire images and videos in situations where acquisition is infeasible for cameras at the ground level [26], [30]–[32]. For this reason, it seems as though human detection technologies deployed through UAVs are bound to redefine the future of various functions.

The topic of human activity recognition (HAR) is not new but the challenges that accompany HAR have not been addressed completely. Several datasets have been used in HAR applications, but these datasets have not addressed challenges that accompany the UCF-ARG aerial dataset [107], [108] that was used in this paper. Traditional datasets have frames that were captured by static ground cameras that show humans in medium-sized or large regions of these frames. Therefore, existing research works propose solutions to detect humans only when the size or scale of people in the frames is medium or large. We still need to study the robustness of existing human detection methods in aerial surveillance to localize humans that are small in scale, i.e., when the distance between the humans and the camera is large. Furthermore, the frames were captured by a moving (i.e., not fixed) aerial camera, such as that of a UAV.

Real-time detection of humans in aerial captured video sequences has certainly not been without challenges. For instance, the size of a human detected on aerial captured video can vary with the altitude of the UAV. The natural variation in the size of humans to be detected also poses a challenge to the technology [8]. Other problems that arise during the acquisition of aerial captured video include dynamical events such as changes in illumination and pronounced degrees of motion blur that result from camera jitter [33], [34]. All these problems need to be addressed to develop a highly robust classification method that is able to distinguish humans from non-humans [8].

Convolutional neural networks (CNNs) have provided machines the ability to use deep learning (DL) in order to detect objects of various sorts. Among the most widely used algorithms for these tasks are YOLOv5 [35], R-CNN [36], Fast R-CNN, Faster R-CNN, Mask R-CNN [37], R-FCN, SqueezeDet [38], EfficientDet [39], MobileNetV2 [40], RetinaNet [76], ShuffleNet [77], and PeleeNet [78]. It is essential that a detection algorithm is able to yield favorable performance metrics at high inference speeds. Several pieces of literature that describe human detection algorithms report such levels of performance using a variety of methods, like body detection [43, 44], head detection [43]–[46], shoulder detection [47], [48], and abundant others [8], [49]–[53]. However, in many instances, solely detecting features associated with human objects do not suffice, as human detection algorithms need to be able to accurately detect them despite the existence of obstructions within the field of view of the camera.

Advances in machine learning (ML) have changed the course of various domains over the past few decades. The use of neural networks in ML has led to considerable advancements in many applications, such as computer vision and natural language processing. As the use of ML for the detection of static objects has reached near-perfect levels of accuracy and precision, researchers have begun venturing into developing relatively newer methods to perform tasks of greater complexity. For instance, we have gone from the simple task of detecting humans in aerial captured videos to the objective of this paper, which is human activity recognition (HAR). The investigation of DL techniques for HAR enables us to extract considerably more meaningful information from digitized data that will significantly enhance various real-world functions, such as those earlier mentioned.

Methods for HAR belong to one of three main categories, namely (1) vision-based, (2) non-vision or sensor-based, and (3) multimodal [54]. Vision-based HAR methods make use of depth cameras to obtain color videos with depth information and acquire information on human movements for recognition [55]. However, these methods are highly susceptible to error that may be caused by variations in illumination from the environment and short range of detection. Non-vision or sensor-based HAR methods, on the other hand, utilize various sensors, such as wearable devices and ambient sensors, or combinations of these, that enable us to acquire information on human movements for purposes of recognition. Combining sensor types makes hybrid sensors that enhance data features to be collected. Doing so enables us to gather sensory information from the real environment, e.g. from cyber-physical-social systems [56]. Magnetic sensors built into smartphones can also readily obtain the position of their users [57]. However, relying on sensor-derived data alone may be challenging because hardware may prove to be costly and privacy concerns prevent large amounts of data from being made public. Moreover, sensor-derived data may need significant domain-specific expertise to obtain appropriate features that an ML model can process and learn from. A high-dimensional and noisy continuous sequence of observations are produced from smartphone sensors. A combination of hierarchical and kernel extreme learning machine (HK-ELM) models were demonstrated to learn features and classify activities [79]. They utilized a feature fusion approach to combine hand-crafted features and H-ELM based learned features. Finally, multimodal HAR techniques allow us to make use of both vision-based and sensor-based data simultaneously to recognize human activity [80]. This is done so that one modality can provide complementary information in order to overcome the limitations of the other modality.

Raw data acquired from sensors and video data from cameras can now be automatically processed and learned by state-of-the art DL techniques, specifically CNNs and recurrent neural networks (RNNs), which have seen

tremendous improvements in performance over the years. For instance, [58] developed a system that considers underlying force patterns derived from first- and second-order dynamics in the input data to classify human actions. They employ a three-layer neural network architecture model, which consists of hierarchical self-organizing maps (SOM) and a supervised neural network model classifier. The SOM in the first layer reduces the dimensionality of the input data and activity patterns in input sequences are extracted to represent posture frames. Moreover, the second layer also consists of an SOM which receives superimposed activity features from the first layer. Because of this, the temporal invariance of the system is ensured. Finally, the clusters in the second layer SOM are labelled using a supervised neural network in the last layer.

Another approach to HAR that is worth discussing is the discrimination of action by observing the coordinates of the joints on a three-dimensional (3D) human skeleton dataset. Over the years, numerous attempts have been made to overcome the challenges of providing an efficient and effective approach to recognizing human activity that leverages these 3D datasets [59]–[70], [112]. Hand-crafted feature techniques entail significant amounts of human intervention in extracting valuable features from skeleton sequences. Moreover, the extraction of localized spatial and temporal information from processed raw skeletal joints for the formulation of DL methods is particularly difficult. In fact, spatio-temporal representations of skeletal sequences, such as DL approaches, are not capable of substantially preserving local and global joint information and often suffer from view dependence, absence of motion, and insufficiency of spatial and temporal information [70]–[73].

In order to address this, [74] proposed a novel method that maps the 3D skeleton joint coordinates into a spatio-temporal image format (STIF). This, in turn, reduces system complexity and provides features that are able to be discriminated better. A system with four main modules, namely spatio-temporal image formation, transfer learning, fusion, and classification was proposed. Here, skeleton joints are converted into STIF which includes spatial and temporal changes for three planes of view. Then images are included in the backbone model comprising three pre-trained networks, namely MobileNetV2, DenseNet121, and ResNet18, each connected to a fully connected layer to extract highly discriminative features. The features extracted from the three planes of view are then fused three different ways. Finally, the fused features are fed into two subsequent fully connected layers to reduce dimensionality before the action is categorized by a softmax classifier.

There is limited research that uses the UCF-ARG aerial dataset [107], [108] because of the following challenges that it comes with:

- An aerial camera mounted onto a payload platform of a 13' Kingfisher^{TM} Aerostat helium balloon,
- Small size of people in human patches for object detection,
- Varying activities, such as raising hands, walking, and bending bodies, performed by people in the human patches.

Human detection and activity recognition (HDAR) using the highly challenging UCF-ARG aerial dataset has been done using various methods [8], [75], [81], [82], [83]. The combination of "The Fastest Pedestrian Detector in the West" (FPDW) [111] and moving object detection was utilized for human detection and tracking in UAV-based videos [81]. Another approach to HAR based on aerial captured video sequences that comprises two phases, namely an offline phase and an inference phase, along with scene stabilization, has also been proposed [8], [75]. The initial phase uses an AlexNet CNN to create a model that classifies between human and nonhuman and another that classifies human activity [75]. The latter phase detects humans and identifies their actions based on models created in the prior phase. Here, HAR is carried out for each frame of the video and for entire sequences of video frames [75]. Because the regions that contained humans were small and the backgrounds contained other objects such as cars, trees, and boxes, the classification method performed poorly with an accuracy of 68%. Recognition of human activities in UAV-based videos from motion features has been explored by using a bag-of-features approach. Here, visual words were utilized to represent motion features, which were described as a frequency count of the words. The SVM classifier served as the activity detector [82]. Lastly, an automated UAV-based DL algorithm consisted of video stabilization using the surf feature selection and Lucas-Kanade method, human area detection using faster R-CNN, and action recognition using a structure combining a three-dimensional CNN architecture and a residual network [83]. To address limitations encountered by methods described in the literature, we propose the use of EfficientDet-D7 which was the top state-of-the-art detector for human detection to improve detection accuracy, and thus, classification accuracy.

This paper has several contributions. Specifically:

- It makes use of a novel HDAR system to detect humans and recognize their activities from aerial captured video sequences.
- To the best of our knowledge, this is the first paper that uses EfficientDet-D7, a state-of-the-art object detector for human detection in videos with dynamical events such as varied altitudes, illumination changes, camera jitter, and variations in viewpoints, object sizes and colors captured from a moving camera attached to an aerial platform.
- It compares and evaluates the performance of three human detectors after adding special distortions on the video frames, such as blur, noise, and illumination changes.
- It includes a comparison between various human detectors, such as YOLOv4 [86], faster R-CNN [99], and EfficientDet [39] in terms of detection accuracy.

- It includes a comparison between various CNN-based feature extractors in terms of accuracy, precision, recall, F1 score, false negative rare (FNR), false positive rate (FPR), and Area Under Curve (AUC).
- It makes use of the highly challenging UCF-ARG dataset for methods evaluation and comparison.
- It explores the concept of cross-domain learning to transfer parameters learned from object detectors and the detection features extracted from the COCO dataset [104] to aerial captured video sequences.
- It explores the concept of cross-domain learning to transfer the parameters learned from pre-trained CNNs and the recognition features extracted from the ImageNet dataset [105] to aerial captured video sequences.

This paper is organized as follows: Section 2 demonstrates the publicly available COCO dataset [106] used for object detection, the publicly available ImageNet dataset [105] used for feature extraction, and the highly challenging UCF-ARG dataset used for human activity recognition. Additionally, state-of-the-art object detection methods such as YOLO4 [86], faster R-CNN [99], and EfficientDet [39] are explored for the purpose of human detection. Furthermore, the use of various pre-trained CNNs such as ResNet50, EfficientnetB0, EfficientnetB4, and EfficientnetB7 for transfer learning and spatial feature extraction is demonstrated. Finally, the use of LSTM for logging temporal features is explained in detail. In Section 3, the experimental setup, and results are discussed to compare between various human detectors and CNN-based feature extractors. Section 4 summarizes the outcome, significance, and plans for future improvements of this work.

II. MATERIALS AND METHODS

This section describes the video datasets utilized in this research work. Furthermore, it discusses various object detection models. Additionally, various convolutional neural networks are demonstrated. Finally, this section explores the recurrent neural network model for the ultimate objective of human activity classification.

A. DATASET OVERVIEW

In this paper, the UCF-ARG dataset [107], [108] that was acquired using three cameras: an Aerial camera, a Rooftop camera, and a Ground camera (ARG) by the University of Central Florida (UCF) was uses. Here, we focused only on the most challenging dataset that contains videos captured by a high-definition aerial camera mounted on the payload platform of a helium balloon with a resolution of 1920×1080 pixels at 60 fps. The challenges in this dataset are summarized as follows: 1) the frames are varied in terms of viewpoints, color of clothing, positions, orientations, and human sizes; 2) the camera altitudes are varied when the airborne platform is moved; 3) the dataset contains ten human activities performed by twelve different individuals and captured from several views.

The environments where the aerial videos were captured include three car parks in various locations. Ten activities, namely digging, throwing, waving, walking, running, clapping, boxing, jogging, carrying, and opening/closing a car trunk were performed four times by each individual. Therefore, 48 videos were recorded for each activity. This paper used five of these activities, namely throwing, waving, digging, walking, and running. Three of these activities were static, i.e., performed in place (waving, dogging throwing), while two were dynamic (walking and running). These five activities were selected to have a fair comparison with other works which used the same five activities for human activity classification [75], [82], [83]. The performance of human detection models and human activity classification models were evaluated using these five activities.

FIGURE 1. Various human samples detected and cropped from video frames. The image data are from [108].

Figure 1 illustrates various human samples detected and cropped from video frames. The human images are of various sizes but were uniformly resized for visualization. The human image patches in Figure 1 have various backgrounds, colors of clothing, human sizes, activities, and viewpoints.

The numbers of frames in the videos were varied between less than ten and a few hundreds. Only ten frames were selected from each video, yielding a total of $240 \times 10 = 2400$ frames for evaluation and comparison between various human detection models. All video frames were considered for the human activity classification task.

B. METHODS

This section discusses various object detection models such as YOLOv4, faster R-CNN, and EfficientDet used for human detection in videos. Additionally, various CNNs such as ResNet and EfficientNet used for feature extraction from video frames are demonstrated. Finally, this section explores the recurrent neural network model called long short-term memory (LSTM) to record the history of extracted features for the ultimate objective of human activity classification.

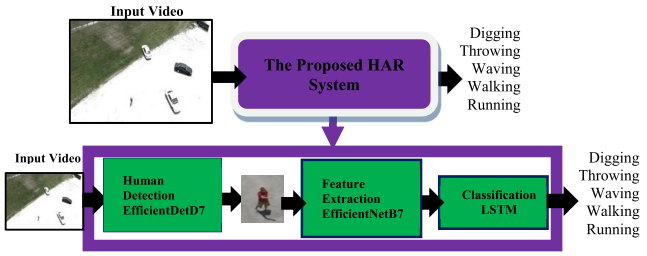

FIGURE 2. The block diagram of the proposed HAR system.

Figure 2 shows the block diagram of the proposed HAR system.

1) HUMAN DETECTION

Various object detectors such as YOLOv4, faster R-CNN, and EfficientDetD7 pre-trained on the COCO dataset with 91 categories of objects were used in this paper for the human detection task. We used these human detectors with learned parameters without finetuning them on the UCF-ARG dataset. In other words, one of our objectives was to evaluate and compare the performance of these object detectors for human detection in challenging aerial videos.

a: YOU ONLY LOOK ONCE (YOLO)

YOLO is a real-time detection model that offers a good balance between high inference speed and good accuracy [84]–[86]. Training in YOLO is performed by applying full images to an end-to-end neural network to utilize convolutional layers for feature extraction and fully connected layers for classification and bounding box prediction. The advantage of YOLO is its ability to see the full images during the training stage which results in a remarkable reduction in the number of background errors when compared with fast R-CNN. However, YOLO has more localization errors. Additionally, YOLO includes 24 convolutional layers added before two fully connected layers. The architecture of YOLO was inspired by the GoogleNet CNN used for image classification. The input image has 224×224 pixels that are fed into convolutional layers for training on the ImageNet dataset. On the other hand, YOLO has double the resolution for detection [84], [85].

YOLOv4 is a more recent, faster and more accurate object detector [86] with many improvements in its architecture and training strategy. It was designed to run in real time and to be trained using only one GPU. YOLOv3 has been demonstrated in various applications and was able to yield high performance in detecting nude humans in pornographic videos [109]. However, it was found that YOLOv4 outperformed YOLOv3 in terms of detection accuracy [86]. YOLOv4 consists of the following components:

1. Backbone module: CSPDarknet53 [87] for feature extraction,
2. Neck module: Spatial Attention Module (SAM) [88], Spatial Pyramid Pooling (SPP) [89], and Path Aggregation Network (PAN) [90] to enhance the receptive field presented by the Backbone
3. Head module: YOLOv3 [91] to predict the final output which are the bounding boxes and the classification scores for each object.

The main improvements in YOLOv4 are in the Neck module and in the training strategy. The Neck module consists of three submodules, namely SPP, modified PAN, and modified SAM. The SPP submodule was added over CSPDarknet53 to increase the receptive field; the modified PAN Net submodule was added as a method of parameter aggregation from different CSPDarknet53 backbone levels; while the modified SAM submodule is an attention mechanism applied over the feature maps.

In this paper, we demonstrated YOLOv4 with the CSPDarknet53 network as a backbone for human detection in aerial videos. The video frames have a size of 540 x 960 pixels that were applied directly to YOLO. YOLOv4 was selected to balance the tradeoff between the accuracy of detection and the speed. YOLOv4 usually runs twice as fast as EfficientDet with comparable performance. In this work, YOLOv4 was tuned to filter and detect only humans and ignore other classes.

b: FASTER REGION BASED CONVOLUTIONAL NEURAL NETWORK (R-CNN)

Region-based CNNs are computationally expensive. Faster R-CNN was found to enhance the detection accuracy and the run time of fast R-CNN. Furthermore, faster R-CNN also outperformed YOLOv3 in terms of detection accuracy [86], [91]. A faster R-CNN [99] object detector consists of two modules. The first module is a region proposal network (RPN) including a fully convolutional network that uses an attention mechanism [100] in order to enhance the feature maps. RPN takes an image of arbitrary size and proposes regions by producing a set of rectangles, each with its objectness score. The feature maps in RPN are shared with the detection network of a fast R-CNN [101], which is the second module that utilizes the proposed regions. To generate region proposals, a small network which has an $n \times n$ spatial window as input is slid over the last shared convolutional feature map and maps the window to lower dimensional features (256-d for ZF). Additionally, the features are fed into two fully connected layers including a box regression layer (reg) and a box classification layer (cls). At each location of the sliding window, k region proposals (boxes) are predicted. The classification layer has $2k$ scores (object or not object for each box), and the regression layer has $4k$ outputs (four coordinates for each box) [99]. Figure 3 illustrates the regional proposal network (RPN).

In this paper, we demonstrated a faster R-CNN with ResNet instead of ZF and VGG for human detection in aerial videos. The video frames have a size of 540×960 pixels that were applied to the detector. The faster R-CNN was tuned to filter and detect only persons and ignore other classes.

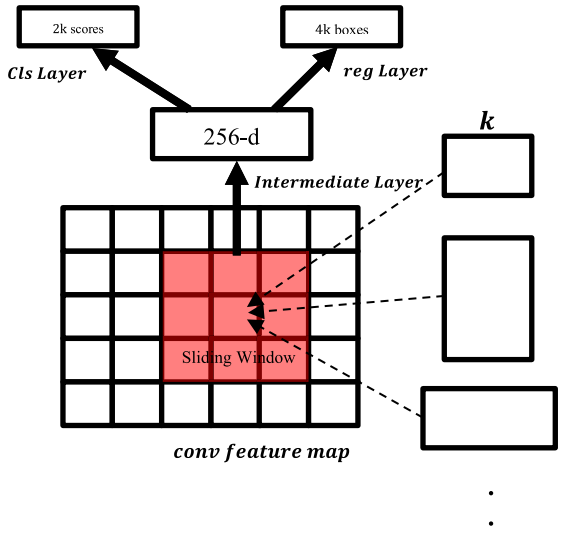

FIGURE 3. Regional Proposal Network (RPN) [99].

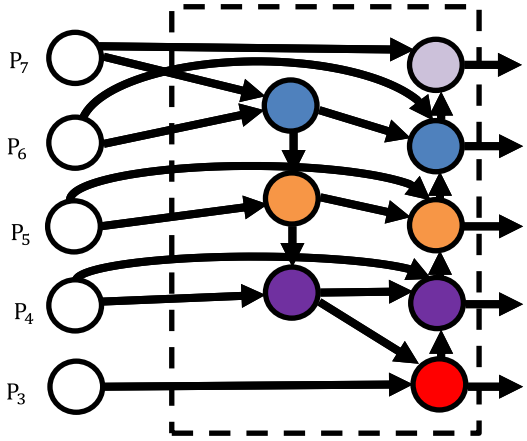

FIGURE 4. The BiFPN features aggregation method [39].

c: EFFICENTDET

EfficientDet is a state-of-the-art object detection method which can yield higher accuracy with much fewer parameters and FLOPs than prior methods [39]. EfficientDet proposed a bi-directional feature pyramid network (BiFPN) used for multi-scale feature fusion. It also has a family of various architectures (D0 ... D7). Additionally, EfficientDet proposed a compound scaling method that uniformly scales the resolution, depth, and width for all backbones, feature networks, and box/class prediction networks simultaneously. Figure 4 shows the BiFPN features aggregation method.

In EfficientDet, BiFPN takes features at levels $\{P_3, P_4, P_5, P_6, P_7\}$ from the EfficientNet [101] backbone. Furthermore, BiFPN is applied repeatedly L times, where L is related to the EfficientDet version. Finally $\{P_3, P_4, P_5, P_6, P_7\}$ are fed into the class/box network. Figure 5 shows the general architecture of the EfficientDet detector.

Where $\emptyset$ is the compound scaling value that is related to the EfficientDet version. Table 1 demonstrates various scaling configurations for EfficientDet (D0 ... D7).

TABLE 1. Scaling configurations for efficientdet (d0 ... d7) [39].

	Input Size	Backbone Network	BiFPN #channels	BiFPN #Layers	Box/Class #Layers
	I_{input}		W_{BiFPN}	D_{BiFPN}	D_{Class}
D0 ($\emptyset = 0$)	512	EfficientNetB0	64	3	3
D1 ($\emptyset = 1$)	640	EfficientNetB1	88	4	3
D2 ($\emptyset = 2$)	768	EfficientNetB2	112	5	3
D3 ($\emptyset = 3$)	896	EfficientNetB3	160	6	4
D4 ($\emptyset = 4$)	1024	EfficientNetB4	224	7	4
D5 ($\emptyset = 5$)	1280	EfficientNetB5	288	7	4
D6 ($\emptyset = 6$)	1280	EfficientNetB6	384	8	5
D7 ($\emptyset = 7$)	1536	EfficientNetB6	384	8	5
D7x	1536	EfficientNetB7	384	8	5

Several experiments were conducted to compare various human detectors including YOLOv4, fast R-CNN, and EfficientDetD7. The outcome of each human detection model is a human patch or Region of Interest (ROI) detected and cropped from each video frame. The comparison was done by selecting ten frames from each of the 240 videos to have 2400 original frames in total. The 2400 selected frames were modified as follows:

1) Flipping frames horizontally,
2) Blurring frames,
3) Adding Gaussian noise to frames,
4) Lightening the frames,
5) Darkening the frames,
6) Converting from the RGB color space to the grayscale color space.

The evaluation and comparison between YOLOv4, fast R-CNN, and EfficientDetD7 was also done on all modified frames. The performance is measured using the detection accuracy as follows:

$$\text{detection accuracy} = \frac{\text{number of frames with correct boxes around humans}}{\text{number of all frames}} \quad (1)$$

C. PRE-TRAINED CNNS FOR FEATURE EXTRACTION

After the human detection model, a set of human patches or ROIs that were cropped from the video frames were applied to CNN-based feature extraction to extract a sequence of features from a sequence of frame ROIs. The sizes of image patches are not equal because altitudes are varied, hence varying the distances between the aerial camera and the human individuals. Therefore, the patches were resized before being applied to pre-trained CNNs to extract the spatial features.

In this paper, the technique of transfer learning was demonstrated to transfer representations from the ImageNet domain

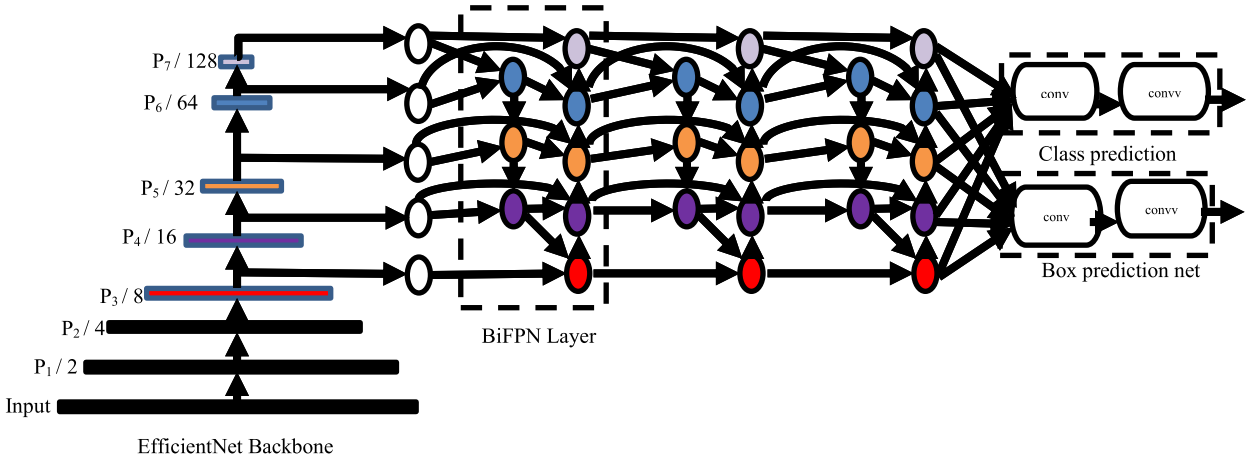

FIGURE 5. The general architecture of the EfficientDet object Detector [39].

to the aerial imagery domain. The parameters of the CNN pre-trained on the ImageNet 1K dataset were utilized for feature extraction without further finetuning of parameters in the first layers. The top layers of pre-trained CNNs were removed and were replaced by LSTM [102], which was tuned with a small-scale UCF-ARG dataset. Various recent architectures of CNNs such as ResNet [103] and EfficientNet [101], including EfficientNetB0, EfficientNetB4, and EfficientNetB7, were pre-trained with natural images of ImageNet to learn high-level (objects and shapes) and low-level (textures, edges, and colors) representations from aerial video frames. The experiments were carried out to compare between the previously mentioned pre-trained CNNs. The image patches or ROIs that were cropped from the EfficientDet human detection model were resized to 224 × 224 in ResNet50, EfficientNetB0, and EfficientNetB4. On the other hand, they were resized to 600 × 600 in EfficientNetB7. The features extracted by each pre-trained CNN have the following dimensions: 2048 in ResNet50, 1280 in EfficientNetB0, 1792 in EfficientNetB4, and 2560 in EfficientNetB7.

d: RESNET

The residual learning framework, also called ResNet, is a very deep network that yields very good generalization without overfitting [103]. In ResNet, different numbers of layers, e.g., 50, 101, and 152 may be used. A supervised learning model feeds a ResNet CNN with large-scale labelled dataset, such as ImageNet in the training stage. The ResNet layers are reformulated as learning residual functions with reference to the layer inputs.

In this paper, the ROIs of humans cropped from aerial video frames were resized to 224 × 224 pixels and were applied to ResNet50 to extract 2048 features. The top layers of ResNet50 were removed. The LSTM was then added to utilize the sequence of features extracted from the sequence of ROIs cropped from the sequence of video frames.

e: EFFICIENTNET

Model scaling is usually done by increasing network depth or network width or by increasing the resolution of input images used for training and evaluation. Even the accuracy is improved through model scaling methods. The drawback, however, is that it entails more manual tuning. While balancing depth, width, and resolution of a network, EfficientNet was found to speed up the inference and outperform the accuracy of existing state-of-the-art CNNs on ImageNet [101]. To improve accuracy, various architectures of EfficientNet are available including B0 as the baseline network; and B1, B2, B3, B4, B5, B6, and B7 as scaling networks. However, more FLOPs is the cost of accuracy improvement.

In this paper, the ROIs of humans cropped from aerial video frames were resized to 224 × 224 pixels in EfficientNetB0 and EfficientNetB4, and to 600 × 600 pixels in EfficientNetB7. Furthermore, they were applied to EfficientNetB0, EfficientNetB4, and EfficientNetB7 to extract 1280, 1792, and 2560 features, respectively. The top layers of EfficientNet were removed. The LSTM was then added to utilize the sequence of features extracted from the sequence of ROIs cropped from the sequence of video frames.

f: LONG SHORT-TERM MEMORY FOR TIME SERIES CLASSIFICATION

A Recurrent Neural Network (RNN) was utilized for sequence modelling to capture temporal correlations [102]. LSTM is a special type of RNN that has been found to slow down gradient vanishing. LSTM has a memory cell to accumulate state information supported by control gates for long-range sequence modelling as shown in Figure 6. In this work, a sequence of features extracted from ROIs of frames were applied to the LSTM that was used to replace the top layers of pre-trained CNNs. Additionally, LSTM was trained, and its parameters were fine-tuned iteratively to fit the features extracted from ROIs cropped from UCF-ARG video frames.

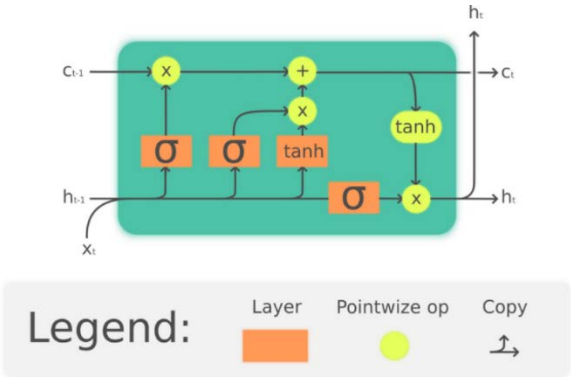

FIGURE 6. An LSTM cell [104].

Several experiments were conducted to select the optimal LSTM structure and architecture that can produce the best performance metrics. The LSTM architecture includes activation functions, the number of LSTM layers, the number of nodes in each layer, and the number of fully connected layers. In this work, the optimal LSTM architecture consists of the following layers:

1) GlobalMaxPool2D
2) Bidirectional_LSTM with 512 nodes, tanh activation and sigmoid recurrent activation
3) Fully connected layers with 256 nodes and ReLU activation function
4) Fully connected layers with five nodes
5) Softmax activation function

Additionally, several experiments were conducted to select the optimal LSTM hyperparameters that can produce the best performance metrics. The hyperparameters include the number of epochs, the optimizer type, the loss function, the learning rate, and the batch size. In this work, the optimal LSTM hyperparameters are as follows:

1) The learning rate used to train the LSTM model was set to 0.001
2) The batch size was set to 32.
3) The number of epochs was set to 50.
4) The loss function was Categorical_Crossentropy.
5) The optimizer was Adam.

In summary, an LSTM architecture with hyperparameters described earlier was used for the classification of a time series including a sequence of features extracted from ROIs cropped from UCF-ARG video frames. The video class or activity class is based on the history of the extracted features. The output of the LSTM is one of five human activities, namely digging, throwing, waving, walking, and running.

Several experiments were conducted to compare various CNN-based feature extraction approaches. The comparison was done by considering all frames in the 240 videos. The performance is measured using the following performance metrics:

1. Accuracy calculates the number of correctly predicted videos over all videos.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

2. Recall (Sensitivity) calculates the number of correctly predicted positive videos over all actual positive videos.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

3. Precision calculates the number of correctly predicted positive videos over all predicted positive videos.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

4. F1 score summarizes recall and precision into one quantity.

$$\text{F1score} = \frac{2 \times precision \times recall}{precision + recall} \tag{5}$$

5. False positive rate (FPR) or false alarm calculates the number of wrongly predicted positive videos over all actual negative videos

$$FPR = \frac{FP}{FP + TN} \tag{6}$$

6. False negative rate (FNR) calculates the number of wrongly predicted negative videos over all actual positive videos

$$FNR = \frac{FN}{FN + TP} \tag{7}$$

In Equations 2 through 7, *TP* is True Positive, *TN* is True Negative, *FP* is False Positive, and *FN* is False Negative.

7. Area Under ROC Curve (AUC) determines the ability of a classifier to distinguish between classes.

The higher the accuracy, recall, precision, F1 score, and AUC are, the better is the model performance. Moreover, the lower the FNR and FPR, the better the model performance is.

III. RESULTS AND DISCUSSION

A. EXPERIMENTAL SETUP

The experiments for human detection and human activity classification were conducted using Python with TensorFlow and OpenCV on Google Colaboratory using an NVIDIA Tesla K80 GPU with 12 GB of memory.

In this work, 48 videos for each of five activities were used, making 240 videos in total. In the human detection task, 2400 frames were selected from 240 videos. The evaluation and comparison between human detection models including YOLOv4, faster R-CNN, and EfficientDet were done to find the best detector that was able to detect humans in the largest number of frames out of 2400 frames. The best detector should also be robust against various effects applied to video frames such as horizontal flipping, blurring, addition of Gaussian noise, lightening, darkening, and conversion from RGB to grayscale.

In the human activity classification task, all patches or ROIs cropped from all video frames were selected from each of 240 videos. The videos were divided into training data and testing data. We followed the same protocol used in state-of-the-art methods that utilized the same UCF-ARG dataset. The

TABLE 2. A comparison between various human detectors using original video frames.

Accuracy	YOLOv4	Faster R-CNN	EfficientDet D0	EfficientDet D4	EfficientDet D7
Digging (original)	0.949	0.936	0.951	0.972	**0.974**
Throwing (original)	0.936	0.909	0.833	0.949	**0.976**
Waving (original)	**0.955**	0.940	0.806	0.953	0.944
Walking (original)	**0.997**	0.968	0.793	0.986	0.988
Running (original)	0.953	0.946	0.700	0.960	**0.973**
Average	0.958	0.940	0.817	0.964	**0.971**

TABLE 3. A comparison between various human detectors using video frames with gaussian noise.

Accuracy	YOLOv4	Faster R-CNN	EfficientDet D0	EfficientDet D4	EfficientDet D7
Digging (Gaussian)	0.761	0.763	0.446	**0.930**	0.902
Throwing Gaussian)	0.774	0.762	0.315	**0.924**	0.911
Waving (Gaussian)	0.825	0.827	0.402	**0.972**	0.944
Walking (Gaussian)	0.906	0.92	0.32	**0.975**	**0.975**
Running (Gaussian)	0.738	0.787	0.215	**0.891**	0.869
Average	0.801	0.812	0.340	**0.938**	0.920

protocol states that one of twelve persons was used for testing and all other eleven persons were used for training and validation. In other words, in each of 12 experiments, the videos were divided into 220 for training and 20 for testing. The evaluation and comparison between pre-trained CNN-based feature extractors were done on ResNet50, EfficientNetB0, EfficientNetB4, and EfficientNetB7. The objective is to find the best pre-trained CNN that would be able to produce the best performance metrics in terms of accuracy, recall, precision, F1 score, False Negative Rate (FNR), False Positive Rate (FPR), and Area Under Curve (AUC).

B. EXPERIMENTS AND RESULTS

1) HUMAN DETECTION EXPERIMENTS

The first set of experiments were conducted to compare between various human detectors including YOLOv4, faster R-CNN, EfficientDetD0, EfficientDetD4, and EfficientDetD7 using video frames. In these experiments, 2400 frames were selected from each set that included original frames and those augmented by flipping, blurring, addition of Gaussian noise, darkening, whitening, and conversion to grayscale.

First, a comparison was done using frames of the original set that contains five activities: digging, waving, throwing, walking, and running. Table 2 shows the detection accuracy of each activity and the average of all five activities. EfficientDetD7 was found to outperform other human detectors with an average accuracy of 97.1%. On the other hand, EfficientDetD0 produced the lowest average accuracy of 81.7%. It is obvious that EfficientDetD7 outperformed others in detecting the following activities: digging, throwing, and running. However, YOLOv4 yielded better accuracies for waving and walking.

a: ABLATION STUDY

We replaced the model of faster R-CNN that was used by Peng *et al.* 2020 [83] to detect and localize humans in video frames. In our research work, we used EfficientDet instead of faster R-CNN and compared the detection accuracy between both models to evaluate their performance when various challenges are available in frames. We found that our proposed human detector was able to improve the detection accuracy significantly and specifically when there are noise, blur, and illumination changes in the frames. Tables 2, 3, 4, 5, 6, 7, 8 show the comparisons between faster R-CNN and EfficientDet.

Second, a comparison was done using frames that contained the five activities with Gaussian noise added. Table 3 shows the detection accuracy of each activity and the average of all five activities. EfficientDetD4 was found to outperform other human detectors with an average accuracy of 93.8%. It is obvious that EfficientDetD4 outperformed others in all activities. Additionally, EfficientDetD7 produced the second highest average accuracy of 92%. On the other hand, other detectors including YOLOv4 and faster R-CNN yielded an average accuracy that was lower than EfficientDetD4 by >10%. Furthermore, the performance of EfficientDetD0 was the worst at only 34%. In summary, the results indicate that EfficientDetD0 was not robust against the Gaussian noise added to the frames. On the contrary, the accuracies of YOLOv4, and faster R-CNN were degraded significantly compared with the ones without noise. Finally, EfficentDetD4 and D7 are highly robust against the addition of Gaussian noise to the video frames.

Third, a comparison was done using frames that have five activities after having been blurred. Table 4 shows the detection accuracy of each activity and the average of all five activities. Although EfficientDetD7 showed degradation in accuracy, it yielded superior performance compared with other human detectors with an average accuracy of 76.3%. Additionally, EfficientDetD7 outperformed other detectors in all five activities. However, other architectures of EfficientDet such as D0 and D4 produced lower accuracies. Furthermore, other detectors including YOLOv4 and faster R-CNN yielded accuracies that were >20% lower than that of EfficientDetD7. In summary, the results indicate that blurring negatively impacts human object detectors in general. However, EfficientDetD7 was still the most robust human detector despite blurring.

Fourth, a comparison was done using frames that have five activities after being flipped horizontally. Table 5 demonstrates the detection accuracy of each activity and the average of all five activities. EfficientDetD7 yielded the best performance with an average accuracy of 97.3%. Although,

TABLE 4. A comparison between various human detectors using video frames with blurring.

Accuracy	YOLOv4	Faster R-CNN	EfficientDet D0	EfficientDet D4	EfficientDet D7
Digging (blur)	0.473	0.520	0.661	0.646	**0.729**
Throwing (blur)	0.463	0.511	0.480	0.671	**0.774**
Waving (blur)	0.610	0.646	0.725	0.772	**0.891**
Walking (blur)	0.544	0.617	0.415	0.668	**0.813**
Running (blur)	0.496	0.543	0.392	0.536	**0.609**
Average	0.517	0.567	0.535	0.659	**0.763**

TABLE 5. A comparison between various human detectors using video frames flipped horizonatally.

Accuracy	YOLOv4	Faster R-CNN	EfficientDet D0	EfficientDet D4	EfficientDet D7
Digging (flip)	0.955	0.955	0.961	**0.983**	0.978
Throwing (flip)	0.955	0.911	0.858	0.949	**0.972**
Waving (flip)	0.955	0.963	0.876	**0.970**	0.948
Walking (flip)	0.984	0.960	0.757	0.984	**0.991**
Running (flip)	0.955	0.946	0.682	0.946	**0.977**
Average	0.961	0.947	0.827	0.966	**0.973**

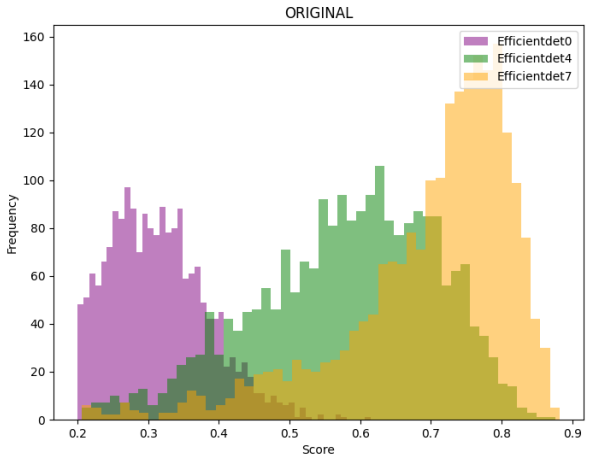

FIGURE 7. Histograms of confidence scores for EfficientDet D0, D4, and D7 using the original video frames.

EfficientDetD7 outperformed other detectors at detecting throwing, walking, and running, EfficientDetD4 was better at detecting digging, and waving. On the other hand, EfficientDetD0 produced the lowest average accuracy of 82.7%.

Fifth, to study the performance of human object detectors in the grayscale color space, a comparison was done using video frames that contained five activities after being converted from RGB to grayscale. Table 6 presents the detection accuracy for each activity and the average of all five activities.

TABLE 6. A comparison between various human detectors using the grayscale version of the video frames.

Accuracy	YOLOv4	Faster R-CNN	EfficientDet D0	EfficientDet D4	EfficientDet D7
Digging (grey)	0.881	0.873	0.775	**0.980**	0.919
Throwing (grey)	0.869	0.852	0.593	**0.943**	0.924
Waving (grey)	0.963	0.917	0.627	**0.968**	0.938
Walking (grey)	**0.986**	0.962	0.557	0.977	0.982
Running (grey)	0.915	0.931	0.521	**0.964**	0.960
Average	0.923	0.907	0.615	**0.966**	0.945

TABLE 7. A comparison between various human detectors after darkening the video frames.

Accuracy	YOLOv4	Faster R-CNN	EfficientDet D0	EfficientDet D4	EfficientDet D7
Digging (dark)	0.957	0.915	0.843	**0.980**	0.970
Throwing (dark)	0.926	0.903	0.665	**0.957**	**0.957**
Waving (dark)	0.955	0.944	0.676	**0.970**	0.959
Walking (dark)	**0.995**	0.964	0.697	0.984	0.988
Running (dark)	0.966	0.966	0.583	0.953	**0.982**
Average	0.960	0.938	0.692	0.969	**0.971**

EfficientDetD4 outperformed other detectors in all activities except walking and produced the best average accuracy of 96.6%. Additionally, EfficientDetD7 yielded the second-best average accuracy. On the other hand, EfficientDetD0 yielded the worst performance with an average accuracy of only 61.5%.

Sixth, to study the performance of human object detectors under various illumination (light) conditions, a comparison was done after darkening and whitening the video frames that contained five activities. Table 7 describes the detection accuracy of each activity and the average of all five activities after darkening the video frames. EfficientDetD7 outperformed other detectors in throwing and running and produced the best average accuracy of 97.1%. Additionally, EfficientDetD7 outperformed other detectors for digging, waving, and throwing and had the second-best average accuracy. Furthermore, YOLOv4 outperformed other detectors for walking. On the other hand, EfficientDetD0 yielded the worst performance with an average accuracy of only 69.2 %.

Table 8 describes the detection accuracy of each activity and the average of all five activities after whitening the video frames. EfficientDetD4 outperformed other detectors for throwing and digging and produced the best average accuracy of 96.5%. Additionally, EfficientDetD7 outperformed other detectors for running and yielded the second-best

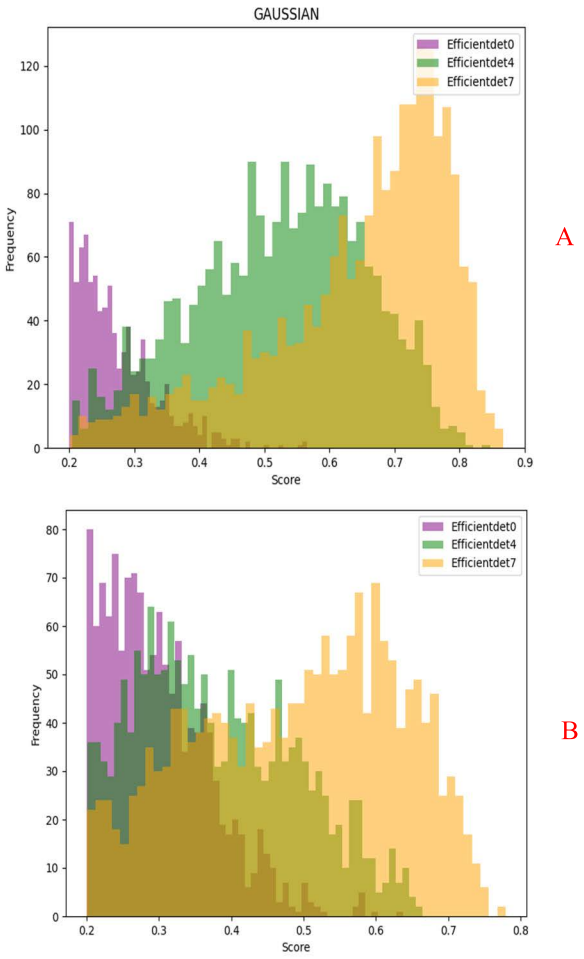

FIGURE 8. Histograms of confidence scores for EfficientDet D0, D4, and D7 using frames A) Gaussian noise, B) blur.

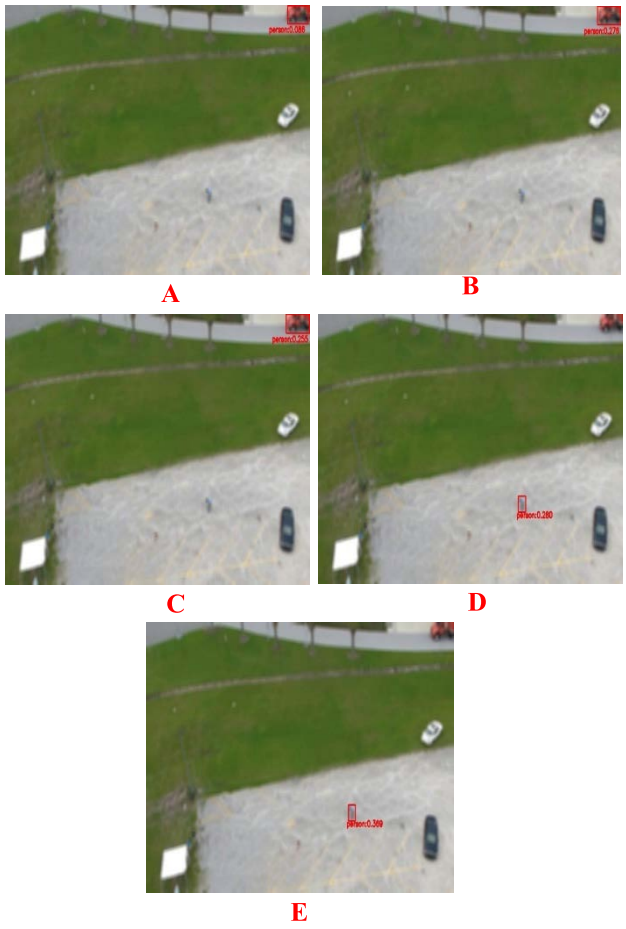

FIGURE 9. Human detection in the first sample frame with blurring using: A.) Faster R-CNN, B) YOLOv4, C) EfficientDetD0, D) EfficientDetD4, and E) EfficientDetD7. The image data are from [108].

TABLE 8. A comparison between various human detectors after whitening the video frames.

Accuracy	YOLOv4	Faster R-CNN	EfficientDet D0	EfficientDet D4	EfficientDet D7
Digging (light)	0.955	0.926	0.930	**0.983**	0.966
Throwing (light)	0.922	0.877	0.776	**0.972**	0.951
Waving (light)	**0.948**	0.931	0.793	0.942	0.936
Walking (light)	**0.993**	0.951	0.775	0.982	0.977
Running (light)	0.944	0.931	0.698	0.944	**0.964**
Average	0.952	0.923	0.794	**0.965**	0.959

average accuracy. Furthermore, YOLOv4 outperformed other detectors for waving and walking. On the other hand, EfficientDetD0 yielded the worst performance with an average accuracy of 79.4%.

A comparison between three architectures of EfficientDet, namely D0, D4, and D7 was demonstrated by plotting the histograms of confidence scores for each detector for 2400 frames. A confidence score reports the probability of prediction of a human category. Figure 7 illustrates the histogram comparison for original video frames. EfficientDetD7 in yellow yielded higher scores than EfficientDetD4 in green. On the other hand, EfficientDetD0 yielded the lowest scores in purple.

Figure 8 compares the histograms of confidence scores after blurring or adding Gaussian noise to video frames. EfficientDetD7 in yellow yielded higher scores than EfficientDetD4 in green. On the other hand, EfficientDetD0 yielded the lowest scores in purple. It is obvious that confidence scores for the original video frames have higher confidence score values than those with Gaussian noise. On the other hand, blurring video frames leads to a reduction in confidence scores for all human detectors. However, EfficientDetD7 was still the most robust detector despite blurring as confirmed by the histograms of confidence scores shown in Figure 8.

In summary, it is obvious based on Tables 2 through 8 that EfficientDetD0 yielded the worst performance with the lowest accuracies in all scenarios. This may be because of the variety of activities performed by humans, and their small

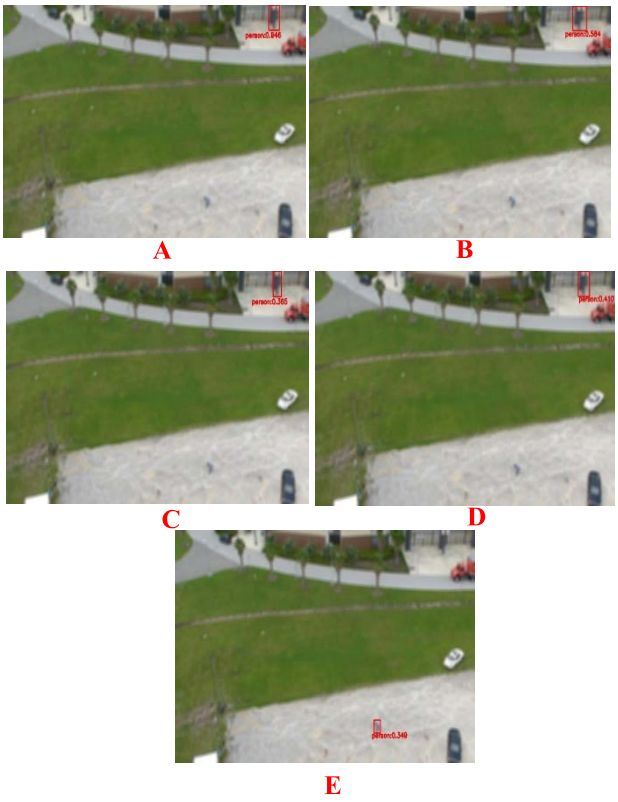

FIGURE 10. Human detection in the second sample frame with blurring using: A.) Faster R-CNN, B) YOLOv4, C) EfficientDetD0, D) EfficientDetD4, and E) EfficientDetD7. The image data are from [108].

sizes in aerial videos. On the other hand, the performances of the YOLOv4 and Faster R-CNN human detectors were good in all scenarios except in those with blurring and Gaussian noise.

Tables 2 through 8 also show that EfficientDetD7 outperformed other human detectors in many scenarios including those in original frames and those augmented by flipping horizontally, by blurring, and by darkening. Similarly, EfficiendDetD7 yielded the second-best performance in other scenarios such as those that were augmented by adding Gaussian noise, by whitening, and by converting to grayscale. The comparison between various human detectors was done by calculating an average accuracy for each human detector for all activities and all scenarios as shown in Table 9.

As a result, we deduced that EfficientDetD7 would be a good human detector that can be utilized in aerial captured video sequences. The power of EfficientDetD7 results from its robustness against various human size, cloth color, views, and positions. Moreover, it can detect humans even in the presence of various factors affecting the video frames such as noise, blur, light change, and grayscale color. Therefore, we utilized EfficientDetD7 in human activity classification experiments to detect and crop ROIs of humans from frames. The cropped patches or ROIs were then applied to CNN-based feature extraction models.

Figures 9, 10, and 11 show three frames with blurring for five human detectors including faster R-CNN,

TABLE 9. Average accuracies for all activities and all scenarioys for various human detectors.

Human Detector	Detection Accuracy
YOLOv4	0.867
Faster R-CNN [85]	0.862
EfficientDetD0	0.661
EfficientDetD4	0.918
EfficientDetD7	**0.929**

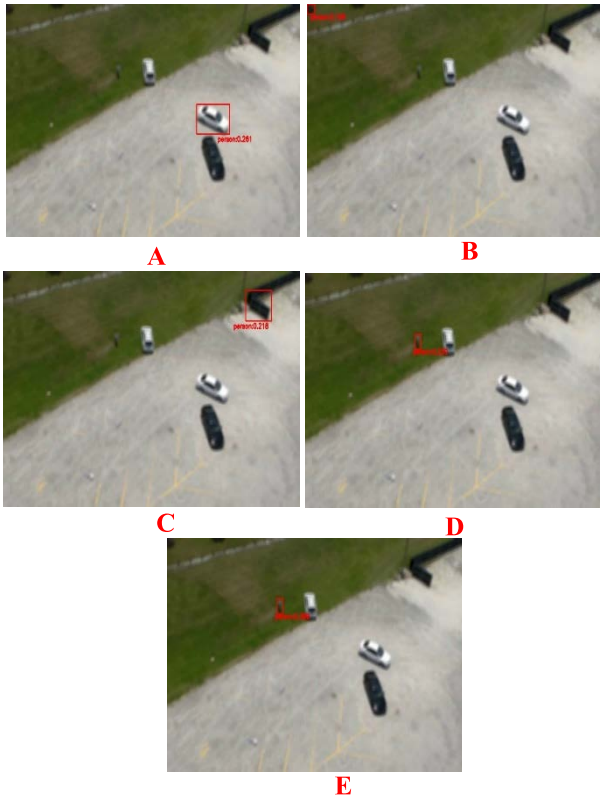

FIGURE 11. Human detection in the third sample frame with blurring, using: A) Faster RCNN, B) YOLOv4, C) EfficientDetD0, D) EfficientDetD4, E) EfficientDetD7. The image data are from [108].

YOLOv4, EfficientDet0, EfficientDetD4, and EfficientDetD7. While EfficientDetD7 was able to detect humans in the three frames, EfficientDet4 was able to detect only two humans in two frames and misclassified one human in one frame. On the other hand, faster R-CNN, YOLOv4, and EfficientDetD0 were not able to detect any humans in all three frames.

They detected only objects that were irrelevant in the background, such as cars.

Figures 12, 13, and 14 show three frames with three scenarios including darkening, converting to grayscale, and whitening, respectively for five human detectors: Faster R-CNN, YOLOv4, EfficientDetD0, EfficientDet4D, and EfficientDetD7. While EfficientDetD7 was able to detect humans in the three frames in the three scenarios, EfficientDetD4 was able to detect only two humans in two frames that were converted to grayscale, and in two frames that were

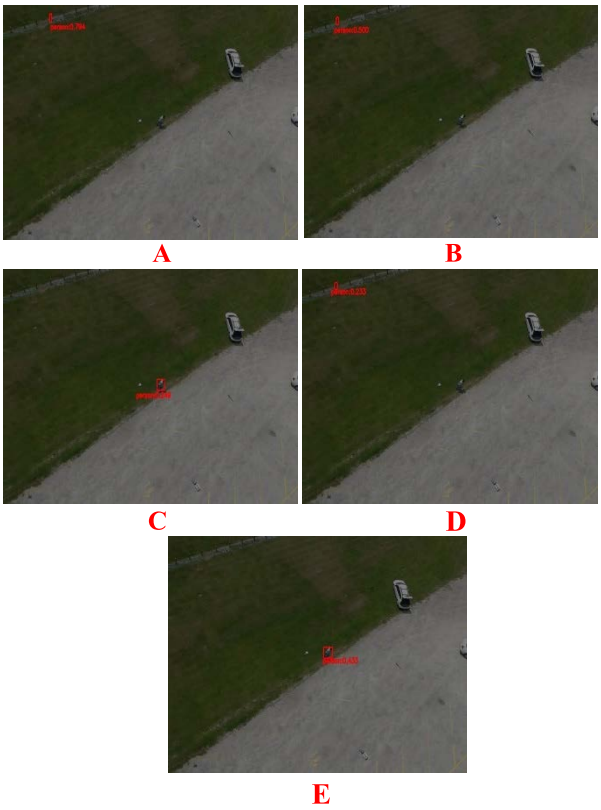

FIGURE 12. Human detection in a sample with darkening using: A) Faster R-CNN, B) YOLOv4, C) EfficientDetD0, D) EfficientDetD4, E) EfficientDetD7. The image data are from [108].

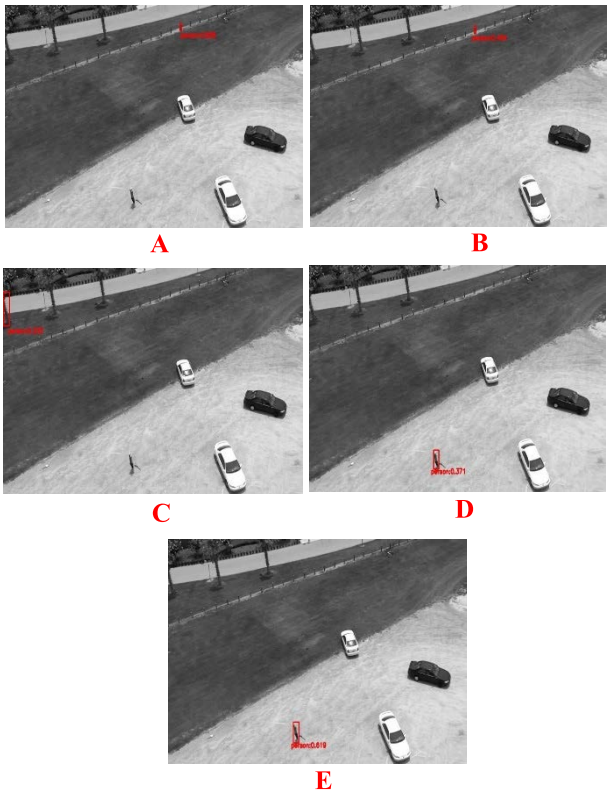

FIGURE 13. Human detection in a sample in the grayscale colour space using: A) Faster R-CNN, B) YOLOv4, C) EfficientDetD0, D) EfficientDetD4, E) EfficientDetD7. The image data are from [108].

whitened. EfficientDetD0 was able to detect only one human in the whitened frames. On the other hand, faster R-CNN and YOLOv4 failed to detect humans in the three frames and detected only objects that were irrelevant in the background.

Figures 15 shows one frame with multiple human individuals detected using five human detectors: Faster R-CNN, YOLOv4, EfficientDetD0, EfficientDet4D, and EfficientDetD7. It is obvious that EfficientDetD7 was able to detect all seven humans in the frame. Furthermore, EfficientDetD4, Faster R-CNN and YOLOv4 were able to detect all humans except one standing behind the white car. On the other hand, EfficientDetD0 was not able to detect five of the humans. It detected only two humans and merged them incorrectly in one box.

Figures 16 shows one frame with multiple humans for five human detectors: Faster R-CNN, YOLOv4, EfficientDetD0, EfficientDet4D, and EfficientDetD7. EfficientDetD7 was able to detect all seven humans in the frame. Furthermore, EfficientDetD4, and Faster R-CNN were able to detect all humans except one standing behind the white car. Additionally, YOLOv4 was able to detect all humans in the frame but also detected irrelevant objects in the background. On the other hand, EfficientDetD0 was not able to detect six of the seven humans.

Figure 17 shows one frame with multiple humans for five human detectors. It is obvious, that EfficientDetD7 was able to detect all eight humans in the frame. While EfficientDetD4, was able to detect seven of the eight humans, it failed to detect the human bending over to pick something up from the ground. Faster R-CNN detected the car trunk as a human and YOLOv4 detected an object on the ground as a human. On the other hand, EfficientDetD0 failed to detect seven of the eight humans on the frame.

Figures 18 shows one frame with multiple humans for five human detectors. It is obvious, that EfficientDetD7 and YOLOv4 were able to detect all eight humans in the frame. However, YOLOv4 detected an object on the ground as a human. While EfficientDetD4, was able to detect seven of the eight humans, it failed to detect the person only whose upper body is visible. Moreover, faster R-CNN was able to detect six of eight humans. On the other hand, EfficientDetD0 failed to detect seven of the eight humans.

Figures 19 shows one frame with multiple humans for five human detectors. It is obvious that EfficientDetD7 was able to detect all seven humans in the frame. On the other hand, EfficientDetD4, Faster R-CNN, and YOLOv4 were able to detect six of the seven humans, but failed to detect the human standing behind the white car. Unfortunately, EfficientDetD0 failed to detect six of the seven humans.

2) HUMAN ACTIVITY CLASSIFICATION EXPERIMENTS

The second set of experiments were conducted to compare between various CNN-based feature extraction models added

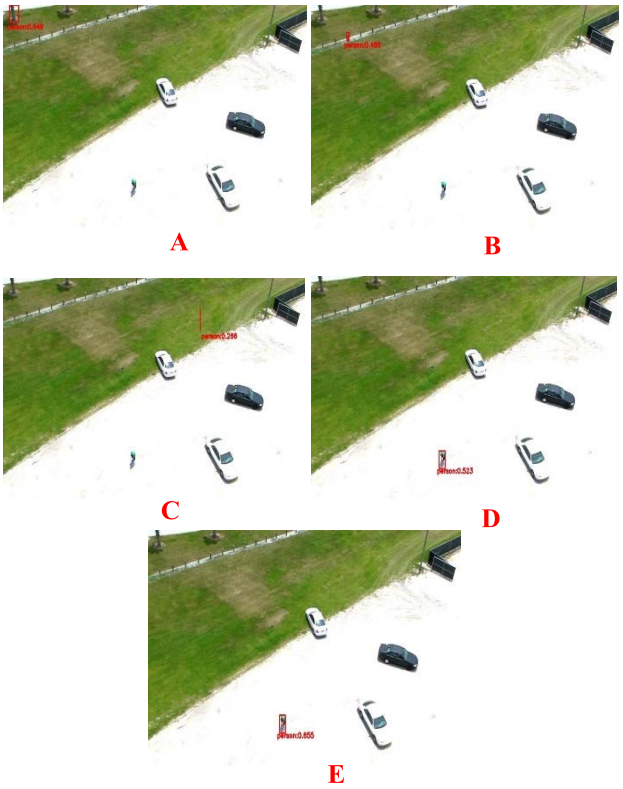

FIGURE 14. Human detection in a sample with whitening using A) Faster R-CNN, B) YOLOv4, C) EfficientDetD0, D) EfficientDetD4, E) EfficientDetD7. The image data are from [108].

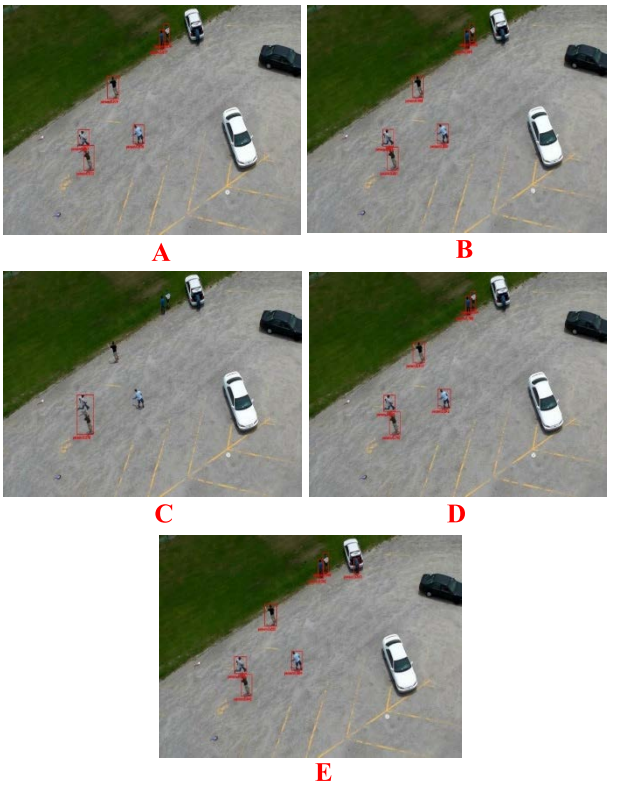

FIGURE 15. First sample frame with multiple humans, A) Yolov4, B) Faster R-CNN, C) Efficientdet0, D) Efficientdet4, E) Efficientdet7. The image data are from [108].

TABLE 10. Classification report of resnet50 for each of 12 persons.

Person	Precision	Recall	F1 Score	Accuracy	FNR	FPR
Person 1	0.81	0.80	0.80	0.80	0.20	0.05
Person 2	0.79	0.75	0.74	0.75	0.25	0.0625
Person 3	0.91	0.85	0.83	0.85	0.15	0.0375
Person 4	0.71	0.70	0.69	0.70	0.30	0.075
Person 5	0.86	0.80	0.80	0.80	0.20	0.05
Person 6	0.93	0.90	0.90	0.90	0.10	0.025
Person 7	0.62	0.55	0.52	0.55	0.45	0.1125
Person 8	0.87	0.80	0.77	0.80	0.20	0.05
Person 9	0.84	0.80	0.79	0.80	0.20	0.05
Person 10	0.82	0.75	0.72	0.75	0.25	0.0625
Person 11	0.71	0.65	0.62	0.65	0.35	0.0875
Person 12	0.77	0.75	0.75	0.75	0.25	0.0625
Average	**0.80**	**0.75**	**0.74**	**0.75**	**0.24**	**0.06**

TABLE 11. Classification report of efficientnetb0 for each of 12 persons.

Person	Precision	Recall	F1 Score	Accuracy	FNR	FPR
Person 1	0.68	0.65	0.63	0.65	0.35	0.0875
Person 2	0.62	0.55	0.51	0.55	0.45	0.1125
Person 3	0.67	0.70	0.68	0.70	0.30	0.075
Person 4	0.72	0.70	0.68	0.70	0.30	0.075
Person 5	0.88	0.70	0.70	0.70	0.30	0.075
Person 6	0.58	0.65	0.58	0.65	0.35	0.0875
Person 7	0.49	0.55	0.51	0.55	0.45	0.1125
Person 8	0.80	0.60	0.59	0.60	0.40	0.1
Person 9	0.89	0.85	0.84	0.85	0.15	0.0375
Person 10	0.77	0.65	0.63	0.65	0.35	0.0875
Person 11	0.48	0.55	0.50	0.55	0.45	0.1125
Person 12	0.72	0.70	0.67	0.70	0.30	0.075
Average	**0.69**	**0.65**	**0.62**	**0.65**	**0.34**	**0.086**

before the LSTM architecture for human activity classification using human ROIs cropped from the video frames. In these experiments, all original video frames were selected from each video. The number of videos for the five activities, namely digging, waving, throwing, walking, and running is 240 (48 for each activity). Tables 10 through 13 show the performance metrics for each CNN, namely ResNet50, EfficientNetB0, EfficientNetB4, and EfficientNetB7. The metrics were calculated, running each model 12 times. Each time, we took one person out of 12 for testing and the other 11 for training and validation. The accuracy, recall, precision, F1 score, FNR, and FPR were calculated for each person for five activities. Additionally, an average for each metric was calculated to compare the CNNs and to find the best candidate for the HAR task.

In Table 10, ResNet50 was evaluated for 12 persons with five activities. An average accuracy of 75% and an average F1 score of 74% were calculated.

In Table 11, the performance of EfficientNetB0 was evaluated for 12 persons with five activities. An average accuracy of 65% and an average F1 score of 62% were calculated. On the other hand, Table 12 demonstrates the performance of EfficientNetB4 for 12 persons with five activities. The average accuracy of EfficientNetB4 was 71% which outperforms

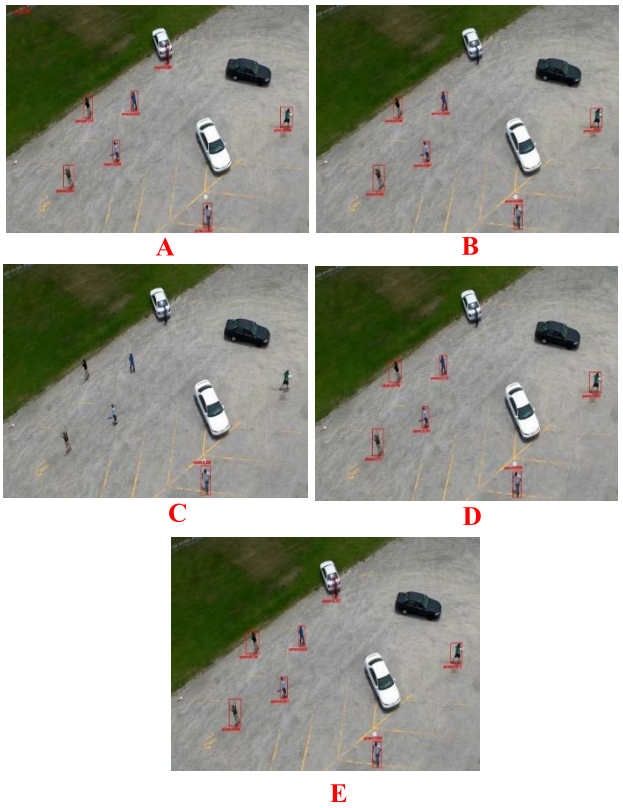

FIGURE 16. Second sample frame with multiple humans, A) Yolov4, B) Faster R-CNN, C) Efficientdet0, D) Efficientdet4, E) Efficientdet7. The image data are from [108].

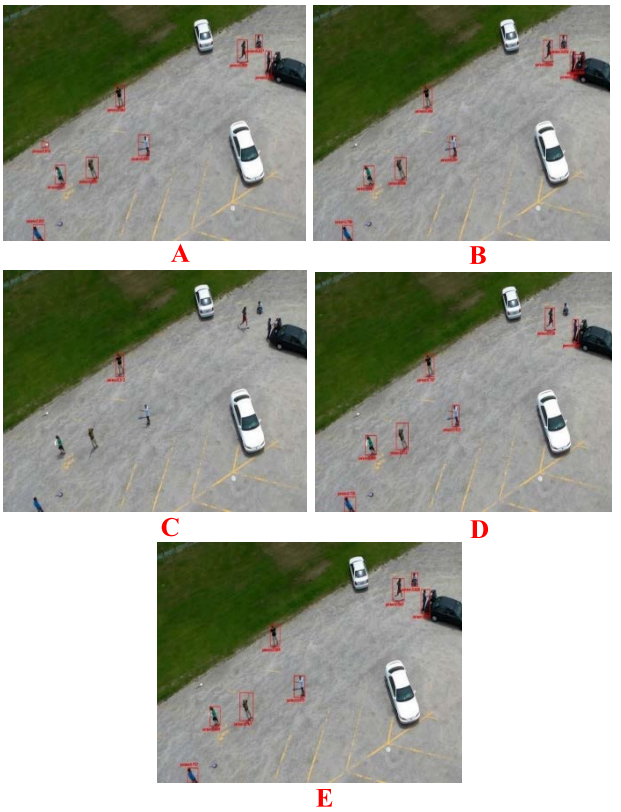

FIGURE 17. Third sample frame with multiple humans, A) Yolov4, B) Faster R-CNN, C) Efficientdet0, D) Efficientdet4, E) Efficientdet7. The image data are from [108].

TABLE 12. Classification report of efficientnetb4 for each of 12 persons.

Person	Precision	Recall	F1 Score	Accuracy	FNR	FPR
Person 1	0.80	0.80	0.80	0.80	0.20	0.05
Person 2	0.59	0.55	0.47	0.55	0.45	0.1125
Person 3	0.86	0.85	0.85	0.85	0.15	0.0375
Person 4	0.77	0.75	0.74	0.75	0.25	0.0625
Person 5	0.85	0.75	0.76	0.75	0.25	0.0625
Person 6	0.82	0.75	0.74	0.75	0.25	0.0625
Person 7	0.49	0.60	0.53	0.60	0.40	0.1
Person 8	0.87	0.80	0.77	0.80	0.20	0.05
Person 9	0.47	0.60	0.51	0.60	0.40	0.1
Person 10	0.81	0.75	0.75	0.75	0.25	0.0625
Person 11	0.73	0.65	0.64	0.65	0.35	0.0875
Person 12	0.78	0.70	0.70	0.70	0.30	0.075
Average	**0.73**	**0.71**	**0.68**	**0.71**	**0.28**	**0.071**

TABLE 13. Classification report of efficientnetb7 for each of 12 persons.

Person	Precision	Recall	F1 Score	Accuracy	FNR	FPR
Person 1	0.89	0.85	0.84	0.85	0.15	0.0375
Person 2	0.73	0.70	0.69	0.70	0.30	0.075
Person 3	0.85	0.85	0.85	0.85	0.15	0.0375
Person 4	0.85	0.85	0.85	0.85	0.15	0.0375
Person 5	0.91	0.85	0.85	0.85	0.15	0.0375
Person 6	0.93	0.90	0.89	0.90	0.10	0.025
Person 7	0.60	0.60	0.57	0.60	0.40	0.1
Person 8	0.92	0.90	0.90	0.90	0.10	0.025
Person 9	0.79	0.75	0.74	0.75	0.25	0.0625
Person 10	0.78	0.75	0.74	0.75	0.25	0.0625
Person 11	0.82	0.80	0.79	0.80	0.20	0.05
Person 12	0.87	0.85	0.84	0.85	0.15	0.0375
Average	**0.828**	**0.804**	**0.795**	**0.80**	**0.20**	**0.048**

the accuracy of EfficientNetB0 by 6%. Additionally, the average F1 score of EfficientNetB4 was 68%, which outperforms F1 score of EfficientNetB0 by 6%. Neither EfficientNetB0 nor EfficentNetB4 was able to outperform ResNet50, which was better by 4% in terms of accuracy and 6% in terms of F1 score.

Table 13 demonstrates the performance of EfficientNetB7 for 12 persons with five activities. An average accuracy of 80% and an average F1 score of 79.5% were calculated. Table 13 shows that EfficientNetB7 outperformed ResNet50 by 5% in terms of accuracy and 5.5% in terms of F1 score.

In summary, EfficientNetB7 was found to outperform all other CNN-based feature extraction models that were used before the LSTM architecture to extract spatial features from ROIs cropped from the video frames. The superior performance of EfficientNetB7 was obvious because it achieved the highest values for the following metrics: accuracy (80%), recall (80.4%), precision (82.8%), and F1 score (79.5%), and the lowest values of FNR (20%), and FPR (4.8%). These results proved that EfficientNetB7 was the best of the candidate CNN-based feature extractors for the HAR task. Figure 20 shows the average for each metric: accuracy,

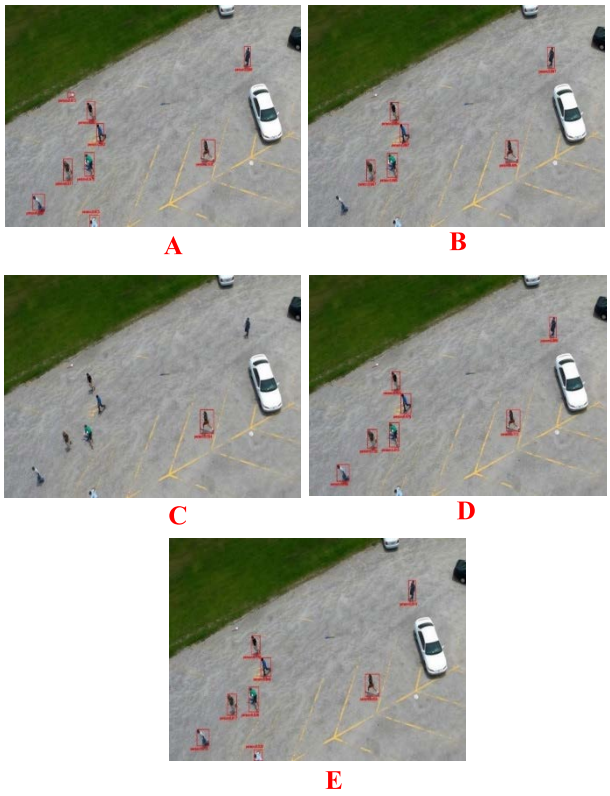

FIGURE 18. Fourth sample frame with multiple humans, A) Yolov4, B) Faster R-CNN, C) Efficientdet0, D) Efficientdet4, E) Efficientdet7. The image data are from [108].

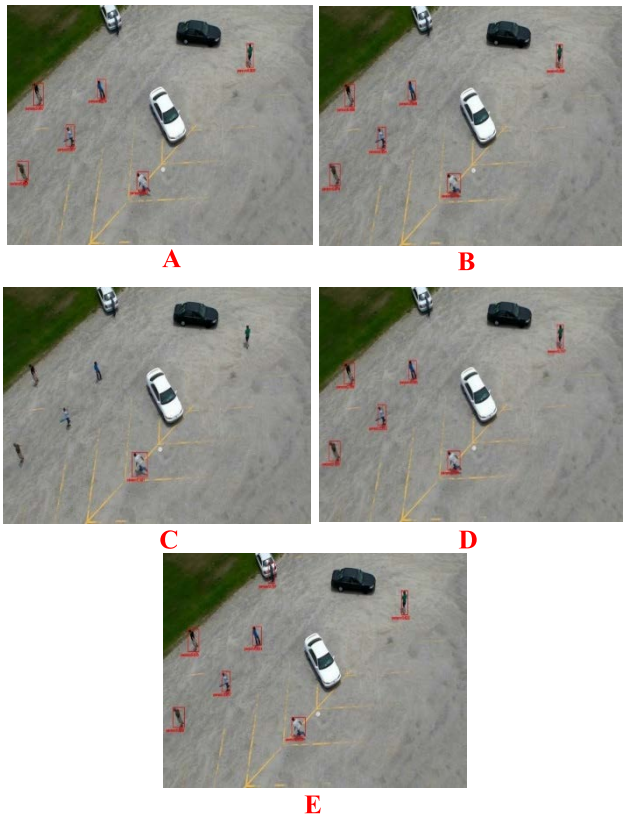

FIGURE 19. Fifth sample frame with multiple humans, A) Yolov4, B) Faster R-CNN, C) Efficientdet0, D) Efficientdet4, E) Efficientdet7. The image data are from [108].

recall, precision, F1 score, FNR, and FPR for the four CNNs combined with the LSTM architecture.

Figure 21 shows twelve confusion matrices. Each matrix was found for each of 12 persons using a combination of EfficientNetB7 and the LSTM architecture. The numbers in the main diagonal are greater than surrounding values. Figure 21 demonstrates the problem of recognizing digging as throwing and throwing as digging in some cases. Additionally, another recognition error appeared between throwing and waving in a few cases. While digging was recognized correctly in nine of 12 cases, throwing and walking were recognized correctly in only five cases. On the other hand, running was the most correctly recognized activity, as it was recognized correctly in most (11 of 12) cases.

The performance of CNN-based feature extraction models was also evaluated in terms of area under curve (AUC). EfficientNetB7+LSTM was found to outperform other CNNs, as it yielded an AUC of 94%. The second top AUC of 93% was achieved by ResNet50. Additionally, EfficientNetB4 obtained the third top AUC of 92%. On the other hand, the worst AUC of 90% was obtained by EfficientNetB0.

Finally, the proposed method that includes EfficientDetD7 for detection, EfficientNetB7 for feature extraction, and LSTM for classification was compared with state-of-the-art methods that used the UCF-ARG dataset for human activity recognition as shown in Table 14. The comparison was

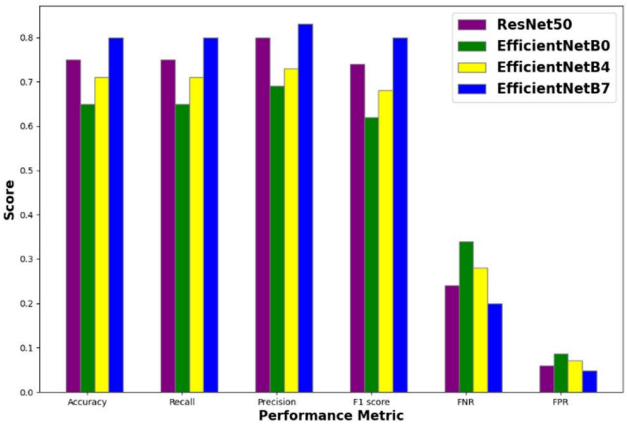

FIGURE 20. Performance metrics of various CNN-based feature extractors combined with LSTM.

done in terms of activity classification accuracy. Burghouts et al [82] proposed the first HAR method that extracted motion features from videos and utilized these features for activity classification. This method yielded an accuracy of 57%. Additionally, Burghouts et al [82] proposed a second method that utilized tracking and focus attention for classification. This method yielded an accuracy of 75%. Furthermore, Hazar et al [75] utilized optical flow stabilization to propose ROIs

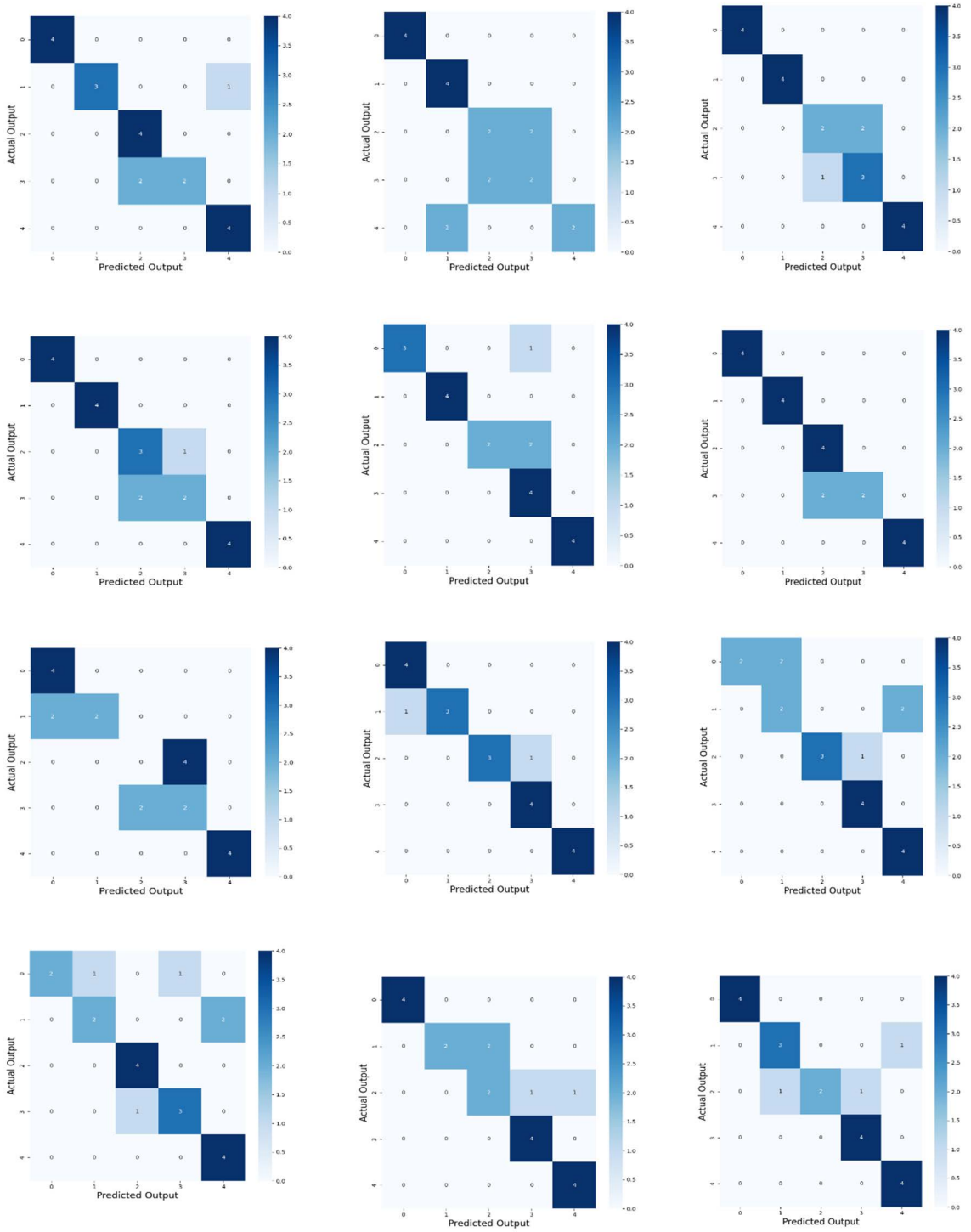

FIGURE 21. Confusion matrix for each of twelve human individuals using EfficientNet7.

that can detect humans using AlexNet and classify activity using GoogleNet. The method yielded an accuracy of 68%. Peng et al [83] also targeted this dataset using speeded-up robust features (SURF) for stabilization, faster R-CNN for detection, and Inception-ResNet-3D for classification. This method yielded an accuracy of 73.72 %. The proposed method was found to outperform other methods by producing an accuracy of 80%.

In summary, deep CNNs pre-trained on ImageNet 1K were used to transfer representations and features from ImageNet to aerial video frames. It was found that EfficientNetB7 representations are more informative when distinguishing

TABLE 14. Average accuracies for all activities and all 12 human individuals for various human activity classification models.

Model	Accuracy
(Burghouts et al. 2014) [84] Motion features without tracking	57 %
(Burghouts et al. 2014) [84] (tracking + focus attention)	75 %
(Peng et al. 2020) [85] SURF (stabilization) + faster R-CNN (detection) + Inception-ResNet-3D (classification)	73.72%
(Hazar et al. 2020) [77] Optical flow (Stabilization) + AlexNet (detection) + GoogleNet (classification)	68 %
EfficientDetD7 (Detection)+EfficientNetB7_LSTM (classification) (proposed)	**80 %**

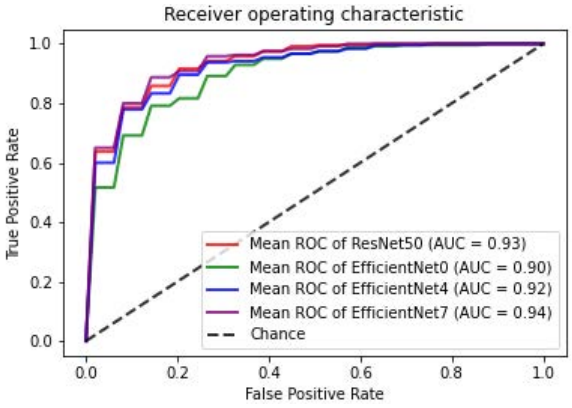

FIGURE 22. Confusion matrix for each of twelve human individuals using EfficientNet7.

between various human activities than other CNNs such as EfficientNetB0, EfficientNetB4, and ResNet50.

The advantages of the proposed HDAR system are as follows:

1. First, the task is formulated as an object detection task. It focuses the attention of the model on human regions inside the video frames and ignores irrelevant objects in the background, such as cars. The outcomes are image patches that include humans detected by EfficientDetD7 and cropped from the frames.
2. Second, the task is formulated as a human activity classification task. Deep CNNs were used to extract spatial features from the cropped human patches. Additionally, an LSTM architecture was utilized to classify a time series of activities into five classes, namely digging, waving, throwing, walking, and running.
3. The proposed system, which combines EfficieneDetD7 for human detection, EfficieneNetB7 for feature extraction, and LSTM for time series classification, is robust against:
   a. various viewpoints, human sizes, and clothing colors.
   b. varied altitudes, illumination changes, and camera jitter.
   c. various conditions, such as blurring, addition of Gaussian noise, lightening, darkening, and conversion from the RGB color space to the grayscale color space.

IV. CONCLUSION AND FUTURE WORK

The work presented in this paper targeted two tasks: human detection (HD) and human activity recognition (HAR). The publicly available UCF-ARG aerial dataset was used to evaluate the performance of the proposed HDAR system. In this video dataset, a moving camera attached to an aerial platform was utilized to capture aerial video sequences. This dataset has highly challenging content with dynamical events such as varied altitudes, illumination changes, camera jitter, and variations in viewpoints, object sizes and colors. Various human object detectors pre-trained on the COCO dataset, such as YOLO, faster R-CNN, and EfficientDet were evaluated to select the best detector that can detect humans and localize them inside the video frames. Several experiments were conducted to compare previously mentioned human detectors. Additionally, various versions of EfficientDet including D0, D4, and D7 were compared. Furthermore, we demonstrated the capability of object detectors to detect humans performing various actions, such as digging, waving, throwing, walking, and running. Second, we added various effects on video frames by flipping horizontally, blurring, adding Gaussian noise, lightening, darkening, and converting RGB to grayscale in order to validate the robustness of the object detectors. The objective of human detection was to detect and crop human patches (ROIs) from video frames. It was found that EfficientDetD7 outperformed other detectors with an average detection accuracy of 92.9%.

This research proposed new challenges to the UCF-ARG aerial dataset by adding various distortions such as blur, noise, and illumination changes. The performance of three human detectors in these poor frame conditions was evaluated. Our evaluation showed that the performance of a faster R-CNN human detector is degraded when these distortions are added. On the other hand, it showed that EfficientDet was robust against these distortions and can detect humans in all conditions included in the evaluation.

Furthermore, several experiments were carried out to compare various deep pre-trained CNNs, such as ResNet50, EfficientNetB0, EfficientNetB4, and EfficientNetB7, which were used to extract spatial features. The extracted features were utilized by LSTM to consider temporal relations between features for human activity classification. Experimental results found that EfficientNetB7-LSTM was able to outperform other CNNs in terms of average accuracy (80%), average precision (83%), average recall of (80%), average F1 score (80%), average false negative rate (FNR) (20%), average

false positive rate (FPR) (4.8%), and average Area Under Curve (AUC) (94%).

The proposed system can be utilized in drones in various industrial applications including surveillance and security, delivery, healthcare and telemedicine, COVID-19 pandemic, and disaster management for human actions surveillance and human behavior understanding.

In summary, a combination of EfficientDetD7 for human detection, EfficientNetB7 for feature extraction, and LSTM for time series classification was proposed to develop a novel HAR system with good performance. The limitation in the proposed HAR system was its poor ability to distinguish between throwing and digging. Moreover, the dataset was small with only 240 videos for five activities. Furthermore, current HAR systems utilize only features extracted from video frames using parameters learned on the ImageNet dataset. In other words, all layers of the CNNs were frozen except the top layers, which were replaced with LSTM. Hence, in the future, we plan to improve the performance of our proposed method by fine-tuning all layers of the CNNs with aerial video frames to enhance accuracy. Lastly, more recently developed deep learning models, such as the vision transformer [110], may be good candidates for the enhancement of recognition performance.

ACKNOWLEDGMENT

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

REFERENCES

[1] M. Paul, S. M. E. Haque, and S. Chakraborty, "Human detection in surveillance videos and its applications—A review," *EURASIP J. Adv. Signal Process.*, vol. 2013, no. 1, pp. 1–16, Dec. 2013, doi: 10.1186/1687-6180-2013-176.

[2] F. Awad and R. Shamroukh, "Human detection by robotic urban search and rescue using image processing and neural networks," *Int. J. Intell. Sci.*, vol. 4, no. 2, pp. 39–53, 2014, doi: 10.4236/ijis.2014.42006.

[3] I. Nourbakhsh, K. Sycara, M. Koes, M. Yong, M. Lewis, and S. Burion, "Human–robot teaming for search and rescue," *IEEE Pervasive Comput.*, vol. 4, no. 1, pp. 72–78, Jan./Mar. 2005, doi: 10.1109/MPRV.2005.13.

[4] P. Doherty and P. Rudol, "A UAV search and rescue scenario with human body detection and geolocalization," in *AI 2007: Advances in Artificial Intelligence*. Berlin, Germany: Springer, 2007, pp. 1–13, doi: 10.1007/978-3-540-76928-6_1.

[5] Z. Uddin and M. Islam, "Search and rescue system for alive human detection by semi-autonomous mobile rescue robot," in *Proc. Int. Conf. Innov. Sci., Eng. Technol. (ICISET)*, Oct. 2016, pp. 1–5, doi: 10.1109/ICISET.2016.7856489.

[6] E. Lygouras, N. Santavas, A. Taitzoglou, K. Tarchanidis, A. Mitropoulos, and A. Gasteratos, "Unsupervised human detection with an embedded vision system on a fully autonomous UAV for search and rescue operations," *Sensors*, vol. 19, no. 16, p. 3542, Aug. 2019, doi: 10.3390/s19163542.

[7] B. Mishra, D. Garg, P. Narang, and V. Mishra, "Drone-surveillance for search and rescue in natural disaster," *Comput. Commun.*, vol. 156, pp. 1–10, Apr. 2020, doi: 10.1016/j.comcom.2020.03.012.

[8] N. AlDahoul, A. Q. M. Sabri, and A. M. Mansoor, "Real-time human detection for aerial captured video sequences via deep models," *Comput. Intell. Neurosci.*, vol. 2018, pp. 1–14, Feb. 2018, doi: 10.1155/2018/1639561.

[9] B. Engberts and E. Gillissen, "Policing from above: Drone use by the police," in *The Future of Drone Use: Opportunities and Threats From Ethical and Legal Perspectives*, B. Custers, Ed. The Hague, The Netherlands: T.M.C. Asser Press, 2016, pp. 93–113, doi: 10.1007/978-94-6265-132-6_5.

[10] R. Stone, T. M. Schnieders, K. A. Push, S. Terry, M. Truong, I. Seshie, and K. Socha, "Human–robot interaction with drones and drone swarms in law enforcement clearing operations," in *Proc. Hum. Factors Ergonom. Soc. Annu. Meeting*, 2019, vol. 63, no. 1, pp. 1213–1217, doi: 10.1177/1071181319631465.

[11] A. Prabhakaran and R. Sharma, "Autonomous intelligent UAV system for criminal pursuit—A proof of concept," *Indian Police J.*, vol. 68, no. 1, pp. 1–20, 2021.

[12] D. Srivastava, S. Shaikh, and P. Shah, "Automatic traffic surveillance system utilizing object detection and image processing," in *Proc. Int. Conf. Comput. Commun. Informat. (ICCCI)*, Jan. 2021, pp. 1–5, doi: 10.1109/ICCCI50826.2021.9402496.

[13] U. Gawande, K. Hajari, and Y. Golhar, "Pedestrian detection and tracking in video surveillance system: Issues, comprehensive review, and challenges," in *Recent Trends in Computational Intelligence*. Rijeka, Croatia: IntechOpen, Jan. 2020, doi: 10.5772/intechopen.90810.

[14] K. Kumar and R. K. Mishra, "A heuristic SVM based pedestrian detection approach employing shape and texture descriptors," *Multimedia Tools Appl.*, vol. 79, nos. 29–30, pp. 21389–21408, Aug. 2020, doi: 10.1007/s11042-020-08864-z.

[15] A. P. M. Singh Gupta, "Video based vehicle and pedestrian detection," *Ann. Romanian Soc. Cell Biol.*, vol. 25, no. 6, pp. 14653–14658, Jun. 2021.

[16] E. Zadobrischi and M. Negru, "Pedestrian detection based on TensorFlow YOLOv3 embedded in a portable system adaptable to vehicles," in *Proc. Int. Conf. Develop. Appl. Syst. (DAS)*, May 2020, pp. 21–26, doi: 10.1109/DAS49615.2020.9108940.

[17] M. Pourhomayoun, "Automatic traffic monitoring and management for pedestrian and cyclist safety using deep learning and artificial intelligence," 2020, doi: 10.31979/mti.2020.1808.

[18] P. Sun and A. Boukerche, "Challenges and potential solutions for designing a practical pedestrian detection framework for supporting autonomous driving," in *Proc. 18th ACM Symp. Mobility Manage. Wireless Access*, New York, NY, USA, Nov. 2020, pp. 75–82, doi: 10.1145/3416012.3424628.

[19] K. Balani, S. Deshpande, R. Nair, and V. Rane, "Human detection for autonomous vehicles," in *Proc. IEEE Int. Transp. Electrific. Conf. (ITEC)*, Aug. 2015, pp. 1–5, doi: 10.1109/ITEC-India.2015.7386891.

[20] Y. Nizam, M. N. H. Mohd, and M. M. A. Jamil, "Human fall detection from depth images using position and velocity of subject," *Proc. Comput. Sci.*, vol. 105, pp. 131–137, Dec. 2017, doi: 10.1016/j.procs.2017.01.191.

[21] S. F. Ali, A. Fatima, N. Nazar, M. Muaz, and F. Idrees, "Human fall detection," in *Proc. INMIC*, Dec. 2013, pp. 101–105, doi: 10.1109/INMIC.2013.6731332.

[22] J. Zhang, C. Wu, and Y. Wang, "Human fall detection based on body posture spatio-temporal evolution," *Sensors*, vol. 20, no. 3, p. 946, Feb. 2020, doi: 10.3390/s20030946.

[23] X. Wang, J. Ellul, and G. Azzopardi, "Elderly fall detection systems: A literature survey," *Frontiers Robot. AI*, vol. 7, 2020. [Online]. Available: 10.3389/frobt.2020.00071.

[24] S. Ezatzadeh, M. R. Keyvanpour, and S. V. Shojaedini, "A human fall detection framework based on multi-camera fusion," *J. Exp. Theor. Artif. Intell.*, pp. 1–20, Jul. 2021, doi: 10.1080/0952813x.2021.1938696.

[25] U. Asif, B. Mashford, S. Von Cavallar, S. Yohanandan, S. Roy, J. Tang, and S. Harrer, "Privacy preserving human fall detection using video data," in *Proc. Mach. Learn. Health NeurIPS Workshop*, vol. 116, Dec. 2020, pp. 39–51. [Online]. Available: https://proceedings.mlr.press/v116/asif20a.html

[26] T. Li, J. Liu, W. Zhang, Y. Ni, W. Wang, and Z. Li, "UAV-human: A large benchmark for human behavior understanding with unmanned aerial vehicles," Aug. 2021, *arXiv:2104.00946*. Accessed: Apr. 28, 2022.

[27] M. Laroze, L. Courtrai, and S. Lefèvre, "Human detection from aerial imagery for automatic counting of shellfish gatherers," in *Proc. 11th Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl. (VISIGRAPP)*, Rome, Italy, 2016, pp. 664–671, doi: 10.5220/0005786506640671.

[28] D. Safadinho, J. Ramos, R. Ribeiro, V. Filipe, J. Barroso, and A. Pereira, "UAV landing using computer vision techniques for human detection," *Sensors*, vol. 20, no. 3, p. 613, Jan. 2020, doi: 10.3390/s20030613.

[29] A. G. Perera, A. Al-Naji, Y. W. Law, and J. Chahl, "Human detection and motion analysis from a quadrotor UAV," in *Proc. IOP Conf., Mater. Sci. Eng.*, vol. 405. Bristol, U.K.: IOP Publishing, Sep. 2018, Art. no. 012003, doi: 10.1088/1757-899x/405/1/012003.

[30] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Computer Vision—ECCV 2016*. Cham, Switzerland, 2016, pp. 445–461, doi: 10.1007/978-3-319-46448-0_27.

[31] F. Nex and F. Remondino, "UAV for 3D mapping applications: A review," *Appl. Geomatics*, vol. 6, no. 1, pp. 1–15, Mar. 2014, doi: 10.1007/s12518-013-0120-x.

[32] C. Torresan, A. Berton, F. Carotenuto, S. F. Di Gennaro, B. Gioli, A. Matese, F. Miglietta, C. Vagnoli, A. Zaldei, and L. Wallace, "Forestry applications of UAVs in Europe: A review," *Int. J. Remote Sens.*, vol. 38, nos. 8–10, pp. 2427–2447, May 2017, doi: 10.1080/01431161.2016.1252477.

[33] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. S. Torr, "Staple: Complementary learners for real-time tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 1401–1409, doi: 10.1109/CVPR.2016.156.

[34] S. Zhang, L. Zhuo, H. Zhang, and J. Li, "Object tracking in unmanned aerial vehicle videos via multifeature discrimination and instance-aware attention network," *Remote Sens.*, vol. 12, no. 16, p. 2646, Aug. 2020, doi: 10.3390/rs12162646.

[35] G. Jocher, "Ultralytics/YOLOv5: V5.0—YOLOv5-P6 1280 models, AWS, Supervise.ly and YouTube integrations," Zenodo, Tech. Rep., 2021, doi: 10.5281/zenodo.4679653.

[36] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," Oct. 2014, *arXiv:1311.2524*. Accessed: Apr. 28, 2022.

[37] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," Jan. 2018, *arXiv:1703.06870*. Accessed: Apr. 28, 2022.

[38] B. Wu, A. Wan, F. Iandola, P. H. Jin, and K. Keutzer, "SqueezeDet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 446–454, doi: 10.1109/CVPRW.2017.60.

[39] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," Jul. 2020, *arXiv:1911.09070*. Accessed: Apr. 29, 2022.

[40] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520, doi: 10.1109/CVPR.2018.00474.

[41] C. Wang, S. Zhao, and R. Zhang, "Occlusion-aware discriminative networks for visual object tracking," in *Proc. 2nd Int. Conf. Robot., Intell. Control Artif. Intell.*, New York, NY, USA, Oct. 2020, pp. 320–326, doi: 10.1145/3438872.3439224.

[42] W. Ouyang and X. Wang, "Joint deep learning for pedestrian detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2056–2063, doi: 10.1109/ICCV.2013.257.

[43] Z. Shao, G. Cheng, J. Ma, Z. Wang, J. Wang, and D. Li, "Real-time and accurate UAV pedestrian detection for social distancing monitoring in COVID-19 pandemic," *IEEE Trans. Multimedia*, vol. 24, pp. 2069–2083, 2022, doi: 10.1109/TMM.2021.3075566.

[44] W. Li, H. Li, Q. Wu, F. Meng, L. Xu, and K. N. Ngan, "HeadNet: An end-to-end adaptive relational network for head detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 2, pp. 482–494, Feb. 2020, doi: 10.1109/TCSVT.2019.2890840.

[45] A. Vora and V. Chilaka, "FCHD: Fast and accurate head detection in crowded scenes," May 2019, *arXiv:1809.08766*. Accessed: Apr. 29, 2022.

[46] Y. Wang, Y. Yin, W. Wu, S. Sun, and X. Wang, "Robust person head detection based on multi-scale representation fusion of deep convolution neural network," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Dec. 2017, pp. 296–301, doi: 10.1109/ROBIO.2017.8324433.

[47] S. Wang, J. Zhang, and Z. Miao, "A new edge feature for head-shoulder detection," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2013, pp. 2822–2826, doi: 10.1109/ICIP.2013.6738581.

[48] C. Zeng and H. Ma, "Robust head-shoulder detection by PCA-based multilevel HOG-LBP detector for people counting," in *Proc. 20th Int. Conf. Pattern Recognit.*, Aug. 2010, pp. 2069–2072, doi: 10.1109/ICPR.2010.509.

[49] M. Ahmad, I. Ahmed, F. A. Khan, F. Qayum, and H. Aljuaid, "Convolutional neural network–based person tracking using overhead views," *Int. J. Distrib. Sensor Netw.*, vol. 16, no. 6, Jun. 2020, Art. no. 1550147720934738, doi: 10.1177/1550147720934738.

[50] E. U. Haq, H. Jianjun, K. Li, and H. U. Haq, "Human detection and tracking with deep convolutional neural networks under the constrained of noise and occluded scenes," *Multimedia Tools Appl.*, vol. 79, nos. 41–42, pp. 30685–30708, Nov. 2020, doi: 10.1007/s11042-020-09579-x.

[51] T. Liu and T. Stathaki, "Faster R-CNN for robust pedestrian detection using semantic segmentation network," *Frontiers Neurorobot.*, vol. 12, p. 64, Oct. 2018, doi: 10.3389/fnbot.2018.00064.

[52] S. Y. Nikouei, Y. Chen, S. Song, R. Xu, B.-Y. Choi, and T. R. Faughnan, "Real-time human detection as an edge service enabled by a lightweight CNN," Apr. 2018, *arXiv:1805.00330*. Accessed: Apr. 29, 2022.

[53] M. K. Vasić and V. Papić, "Multimodel deep learning for person detection in aerial images," *Electronics*, vol. 9, no. 9, p. 1459, Sep. 2020, doi: 10.3390/electronics9091459.

[54] F. Fereidoonian, F. Firouzi, and B. Farahani, "Human activity recognition: From sensors to applications," in *Proc. Int. Conf. Omni-Layer Intell. Syst. (COINS)*, Aug. 2020, pp. 1–8, doi: 10.1109/COINS49042.2020.9191417.

[55] C. Jia, Y. Kong, Z. Ding, and Y. Fu, "RGB-D action recognition," in *Human Activity Recognition and Prediction*, 1st ed., Y. Fu, Ed. Cham, Switzerland: Springer, 2016, pp. 87–106.

[56] W. Yang, X. Liu, L. Zhang, and L. T. Yang, "Big data real-time processing based on storm," in *Proc. 12th IEEE Int. Conf. Trust, Secur. Privacy Comput. Commun.*, Jul. 2013, pp. 1784–1787, doi: 10.1109/TrustCom.2013.247.

[57] I. Ashraf, Y. B. Zikria, S. Hur, A. K. Bashir, T. Alhussain, and Y. Park, "Localizing pedestrians in indoor environments using magnetic field data with term frequency paradigm and deep neural networks," *Int. J. Mach. Learn. Cybern.*, vol. 12, no. 11, pp. 3203–3219, Nov. 2021, doi: 10.1007/s13042-021-01279-8.

[58] Z. Gharaee, P. Gärdenfors, and M. Johnsson, "First and second order dynamics in a hierarchical SOM system for action recognition," *Appl. Soft Comput.*, vol. 59, pp. 574–585, Oct. 2017, doi: 10.1016/j.asoc.2017.06.007.

[59] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 588–595, doi: 10.1109/CVPR.2014.82.

[60] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4570–4579, doi: 10.1109/CVPR.2017.486.

[61] L. Xia, C.-C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 20–27, doi: 10.1109/CVPRW.2012.6239233.

[62] Y. Zhu, W. Chen, and G. Guo, "Fusing spatiotemporal features and joints for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 486–491, doi: 10.1109/CVPRW.2013.78.

[63] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2752–2759, doi: 10.1109/ICCV.2013.342.

[64] Y. Yang, C. Deng, S. Gao, W. Liu, D. Tao, and X. Gao, "Discriminative multi-instance multitask learning for 3D action recognition," *IEEE Trans. Multimedia*, vol. 19, no. 3, pp. 519–529, Mar. 2017, doi: 10.1109/TMM.2016.2626959.

[65] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1110–1118, doi: 10.1109/CVPR.2015.7298714.

[66] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," Apr. 2017, *arXiv:1704.02581*. Accessed: Apr. 29, 2022.

[67] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," Mar. 2016, *arXiv:1603.07772*. Accessed: Apr. 29, 2022.

[68] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," Jul. 2016, *arXiv:1607.07043*. Accessed: Apr. 29, 2022.

[69] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. AAAI Conf. Artif. Intell.*, Feb. 2017, vol. 31, no. 1, Art. no. 1. Accessed: Apr. 28, 2022. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/11212

[70] J. Liu, N. Akhtar, and A. Mian, "Skepxels: Spatio-temporal image representation of human skeleton joints for action recognition," Aug. 2018, *arXiv:1711.05941*. Accessed: Apr. 28, 2022.

[71] J. Ren, N. H. Reyes, A. L. C. Barczak, C. Scogings, and M. Liu, "An investigation of skeleton-based optical flow-guided features for 3D action recognition using a multi-stream CNN model," in *Proc. IEEE 3rd Int. Conf. Image, Vis. Comput. (ICIVC)*, Jun. 2018, pp. 199–203, doi: 10.1109/ICIVC.2018.8492894.
[72] B. Li, M. He, X. Cheng, Y. Chen, and Y. Dai, "Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN," Jun. 2017, *arXiv:1704.05645*. Accessed: Apr. 28, 2022.
[73] C. Li, Y. Hou, P. Wang, and W. Li, "Joint distance maps based action recognition with convolutional neural networks," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 624–628, May 2017, doi: 10.1109/LSP.2017.2678539.
[74] N. Tasnim, M. K. Islam, and J.-H. Baek, "Deep learning based human activity recognition using spatio-temporal image formation of skeleton joints," *Appl. Sci.*, vol. 11, no. 6, p. 2675, Mar. 2021, doi: 10.3390/app11062675.
[75] H. Mliki, F. Bouhlel, and M. Hammami, "Human activity recognition from UAV-captured video sequences," *Pattern Recognit.*, vol. 100, Apr. 2020, Art. no. 107140, doi: 10.1016/j.patcog.2019.107140.
[76] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," Feb. 2018, *arXiv:1708.02002*. Accessed: Apr. 28, 2022.
[77] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," Dec. 2017, *arXiv:1707.01083*. Accessed: Apr. 28, 2022.
[78] R. J. Wang, X. Li, and C. X. Ling, "Pelee: A real-time object detection system on mobile devices," Jan. 2019, *arXiv:1804.06882*. Accessed: Apr. 28, 2022.
[79] N. AlDahoul, R. Akmeliawati, and Z. Zaw, "Feature fusion: H-ELM based learned features and hand-crafted features for human activity recognition," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 7, pp. 509–514, Jul. 2019, doi: 10.14569/IJACSA.2019.0100770.
[80] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions," *Knowl.-Based Syst.*, vol. 223, Jul. 2021, Art. no. 106970, doi: 10.1016/j.knosys.2021.106970.
[81] A. W. M. van Eekeren, J. Dijk, and G. Burghouts, "Detection and tracking of humans from an airborne platform," in *Electro-Optical and Infrared Systems: Technology and Applications XI*, vol. 9249. Bellingham, WA, USA: SPIE, Oct. 2014, pp. 249–255, doi: 10.1117/12.2067568.
[82] G. J. Burghouts, A. W. M. van Eekeren, and J. Dijk, "Focus-of-attention for human activity recognition from UAVs," *Proc. SPIE*, vol. 9249, pp. 256–267, Oct. 2014, doi: 10.1117/12.2067569.
[83] H. Peng and A. Razi, "Fully autonomous UAV-based action recognition system using aerial imagery," in *Advances in Visual Computing*. Cham, Switzerland, 2020, pp. 276–290, doi: 10.1007/978-3-030-64556-4_22.
[84] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," May 2016, *arXiv:1506.02640*. Accessed: Apr. 28, 2022.
[85] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," Dec. 2016, *arXiv:1612.08242*. Accessed: Apr. 28, 2022.
[86] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," Apr. 2020, *arXiv:2004.10934*. Accessed: Apr. 28, 2022.
[87] C.-Y. Wang, H.-Y. M. Liao, I.-H. Yeh, Y.-H. Wu, P.-Y. Chen, and J.-W. Hsieh, "CSPNet: A new backbone that can enhance learning capability of CNN," Nov. 2019, *arXiv:1911.11929*. Accessed: Apr. 28, 2022.
[88] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," Jul. 2018, *arXiv:1807.06521*. Accessed: Apr. 28, 2022.
[89] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," 2014, *arXiv:1406.4729*.
[90] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," Sep. 2018, *arXiv:1803.01534*. Accessed: Apr. 28, 2022.
[91] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," Apr. 2018, *arXiv:1804.02767*. Accessed: Apr. 28, 2022.
[92] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," Nov. 2019, *arXiv:1911.08287*. Accessed: Apr. 28, 2022.
[93] Z. Yao, Y. Cao, S. Zheng, G. Huang, and S. Lin, "Cross-iteration batch normalization," Mar. 2021, *arXiv:2002.05712*. Accessed: Apr. 28, 2022.
[94] G. Ghiasi, T.-Y. Lin, and Q. V. Le, "DropBlock: A regularization method for convolutional networks," Oct. 2018, *arXiv:1810.12890*. Accessed: Apr. 28, 2022.
[95] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," May 2017, *arXiv:1608.03983*. Accessed: Apr. 28, 2022.
[96] D. Misra, "Mish: A self regularized non-monotonic activation function," Aug. 2020, *arXiv:1908.08681*. Accessed: Apr. 28, 2022.
[97] S. Yun, D. Han, S. Joon Oh, S. Chun, J. Choe, and Y. Yoo, "CutMix: Regularization strategy to train strong classifiers with localizable features," Aug. 2019, *arXiv:1905.04899*. Accessed: Apr. 28, 2022.
[98] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," Dec. 2015, *arXiv:1512.00567*. Accessed: Apr. 28, 2022.
[99] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," Jan. 2016, *arXiv:1506.01497*. Accessed: Apr. 28, 2022.
[100] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," Jun. 2015, *arXiv:1506.07503*. Accessed: Apr. 28, 2022.
[101] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," Sep. 2020, *arXiv:1905.11946*. Accessed: Apr. 28, 2022.
[102] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.
[103] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," Dec. 2015, *arXiv:1512.03385*. Accessed: Apr. 28, 2022. [Online]. Available: http://arxiv.org/abs/1512.03385
[104] G. Chevalier, "LARNN: Linear attention recurrent neural network," Aug. 2018, *arXiv:1808.05578*. Accessed: Apr. 28, 2022.
[105] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.
[106] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Computer Vision—ECCV 2014*. Cham, Switzerland, 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.
[107] *New System Performs Persistent Wide-Area Aerial Surveillance*. Accessed: Apr. 28, 2022. [Online]. Available: https://spie.org/news/3006-new-system-performs-persistent-wide-area-aerial-surveillance?ArticleID=x41092&SSO=1
[108] *CRCV | Center for Research in Computer Vision at the University of Central Florida*. Accessed: Apr. 28, 2022. [Online]. Available: https://www.crcv.ucf.edu/data/UCF-ARG.php
[109] N. AlDahoul, H. A. Karim, M. H. L. Abdullah, M. F. A. Fauzi, A. S. B. Wazir, S. Mansor, and J. See, "Transfer detection of YOLO to focus CNN's attention on nude regions for adult content detection," *Symmetry*, vol. 13, no. 1, p. 26, Dec. 2020, doi: 10.3390/sym13010026.
[110] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16 × 16 words: Transformers for image recognition at scale," Jun. 2021, *arXiv:2010.11929*. Accessed: Apr. 28, 2022.
[111] P. Dollar, S. Belongie, and P. Perona, "The fastest pedestrian detector in the west," in *Proc. Brit. Mach. Vis. Conf.*, 2010, pp. 68.1–68.11, doi: 10.5244/C.24.68.
[112] C. Patel, D. Bhatt, U. Sharma, R. Patel, S. Pandya, K. Modi, N. Cholli, A. Patel, U. Bhatt, M. A. Khan, S. Majumdar, M. Zuhair, K. Patel, S. A. Shah, and H. Ghayvat, "DBGC: Dimension-based generic convolution block for object recognition," *Sensors*, vol. 22, no. 5, p. 1780, Feb. 2022.

NOUAR ALDAHOUL received the B.Eng. and M.Eng. degrees in computer engineering from Damascus University, in 2008 and 2012, respectively, and the Ph.D. degree in machine learning from International Islamic University Malaysia, in 2019. She is currently a Researcher with the Faculty of Engineering, Multimedia University, Malaysia. Her main research interests include deep learning, computer vision, and the Internet of Things. She was a recipient of several awards, such as a Special Award in RICES 2021 and multiple awards from Malaysia Technology Expo 2022. She also led the team that won the Championship in a Challenge Session of ICIP 2020.

HEZERUL ABDUL KARIM (Senior Member, IEEE) received the B.Eng. degree in electronics from the University of Wales Swansea, U.K., in 1998, with a focus on communications, the M.Eng. degree in science from Multimedia University, Malaysia, in 2003, and the Ph.D. degree from the University of Surrey, U.K., in 2008. He is currently an Associate Professor with the Faculty of Engineering, Multimedia University. His research interests include telemetry, error resilience and multiple description video coding for 2D/3D image/video coding and transmission, and content-based image/video recognition. He is also serving as a Treasurer for the IEEE Signal Processing Society Malaysia Chapter.

AZNUL QALID MD. SABRI (Senior Member, IEEE) received the master's degree by joining a Research Internship Program from the Commonwealth Scientific Research Organization (CSIRO), Brisbane, QLD, Australia, focusing on medical imaging, and the Ph.D. degree (Hons.) on the topic of human action recognition, under a program jointly offered by a well-known research institution in France, Mines de Douai (a Research Laboratory) and the reputable University of Picardie Jules Verne, Amiens, France. He is currently a Senior Lecturer at the Department of Artificial Intelligence, Faculty of Computer Science and Information Technology (FCSIT), University of Malaya, Malaysia. He is a Graduate of the Prestigious Erasmus Mundus Master's in Vision and Robotics (ViBot) and a Master's Program jointly coordinated by three different universities (University of Burgundy, France; University of Girona, Spain; and Heriot-Watt University Edinburgh, U.K.). He is an Active Researcher in the field of artificial intelligence, having published in multiple international conferences as well as international journals. His main research interests include computer vision, robotics, and machine learning. He is one of the pioneering members of FCSIT's COVIRO (Cognitive, Vision and Robotics) Research Group. He is also the principal investigator of multiple research grants.

MYLES JOSHUA TOLEDO TAN (Member, IEEE) was born in Bacolod, Philippines, in 1996. He received the B.S. degree *(summa cum laude)* in biomedical engineering from the University at Buffalo, The State University of New York, in 2017, and the M.S. degree in applied biomedical engineering from Johns Hopkins University, Baltimore, MD, USA, in 2018. He has been an Assistant Professor of chemical engineering with the University of St. La Salle (USLS), since 2018, where he was appointed as an Assistant Professor of natural sciences, in 2020. He has been actively involved in the education and training of students with the Department of Electronics Engineering and the Department of Electrical Engineering, USLS. He also leads the Tan Medical Image and Signal Processing Group (formerly the Tan Research Group). His research interests include biomedical signal processing, medical imaging, deep learning, and engineering and mathematics education. He is a member of the Institute of Physics, U.K., the Institute of Mathematics and its Applications, U.K., the Association for the Advancement of Artificial Intelligence, and the Tau Beta Pi—The Engineering Honor Society (USA). He was a recipient of the Tau Beta Pi Engineering Honor Society Record Scholarship and the Grace Capen Award.

MHD. ADEL MOMO received the B.Eng. degree in software engineering from Yarmouk Private University, in 2020. He is currently working as a Research and Development Embedded Software Engineer at FMS Tech. His main research interests include deep learning, computer vision, natural language processing, and embedded systems. He was a recipient of several awards, such as the ICIP 2020 Challenge Award.

JAMIE LEDESMA FERMIN (Student Member, IEEE) was born in Bacolod, Philippines, in 1999. He is currently pursuing the bachelor's degree with the Electronics Engineering Program, University of St. La Salle (USLS), Bacolod. He has been an Undergraduate Research Assistant with the Tan Medical Image and Signal Processing Group, USLS, since 2018. His research interests include biomedical signal processing, medical imaging, and deep learning.