Computer Communications 156 (2020) 1–10

Contents lists available at ScienceDirect

Computer Communications

journal homepage: www.elsevier.com/locate/comcom

Drone-surveillance for search and rescue in natural disaster

Balmukund Mishra [a], Deepak Garg [a], Pratik Narang [b], Vipul Mishra [a,*]

[a] Department of Computer Science and Engineering, Bennett University, Greater Noida, India
[b] Department of CSIS, BITS Pilani, Pilani, India

ARTICLE INFO

Keywords:
Drone surveillance
Convolution neural network (CNN)
Object detection (OD)
Action recognition
Aerial action dataset

ABSTRACT

Due to the increasing capability of drones and requirements to monitor remote areas, drone surveillance is becoming popular. In case of natural disaster, it can scan the wide affected-area quickly and make the search and rescue (SAR) faster to save more human lives. However, using autonomous drone for search and rescue is least explored and require attention of researchers to develop efficient algorithms in autonomous drone surveillance. To develop an automated application using recent advancement of deep-learning, dataset is the key. For this, a substantial amount of human detection and action detection dataset is required to train the deep-learning models. As dataset of drone surveillance in SAR is not available in literature, this paper proposes an image dataset for human action detection for SAR. Proposed dataset contains 2000 unique images filtered from 75,000 images. It contains 30000 human instances of different actions. Also, in this paper various experiments are conducted with proposed dataset, publicly available dataset, and stat-of-the art detection method. Our experiments shows that existing models are not adequate for critical applications such as SAR, and that motivates us to propose a model which is inspired by the pyramidal feature extraction of SSD for human detection and action recognition Proposed model achieves 0.98mAP when applied on proposed dataset which is a significant contribution. In addition, proposed model achieve 7% higher mAP value when applied to standard Okutama dataset in comparison with the state-of-the-art detection models in literature.

1. Introduction

Search and rescue (SAR) have been a human-intensive task so far, but recent technological advancements can make it autonomous. Using drone surveillance with a recent computer vision technology can increase the number of humans saved at the time of disaster. However, recent activities of using the drone have mixed reaction, but using a drone to save someone's life is novel and a great cause. Recently, drones are being used more for SAR and providing excellent support in those operations. These days, police and fire departments have also adopted drones and collaborated with local SAR teams for time-sensitive rescue operations. In January of 2019, a SAR team in Snowy Canyon State Park, Utah, used a drone to help rescue a hiker trapped on a ledge at night. The hiker was 60 years old, and SAR first found that he was trapped from other hikers who heard the man calling out for help [1]. Also, a rescue event in Texas, drone was used to find an 88-year-old missing man [1]. In another story of using the drone for SAR where it is used to find two cousins trapped on a mountainside in Iceland [1]. All these incidents show the capability and importance of the drone in the field of SAR however, these operation are performed manually for finding the person which need an automation to apply at bigger scale. These autonomous drone with on-device video analysis capability for saving life motivates us to develop a novel and dedicated system for autonomous searching of people who are stuck and required rescue.

The idea of drone-surveillance for SAR is to use the drone for scanning the affected area with the help of camera, and model deployed on the drone itself for identifying the exact places where help is required. An example of automated surveillance and search operation is shown in Fig. 1. In this figure, after the identification of humans location, GPS location of human can be sent to the rescue team for the fast and productive rescue. The recent success of deep-learning approaches for object detection and action recognition motivates us to apply it in the drone-surveillance. The essential part of a deep-learning approach is that a significant amount of data is needed for training. Since, most of the dataset available in the literature are for ground-level surveillance such as UCF [2], which is not useful for training deep-learning model of aerial surveillance. Hence, it is our primary objective to develop a dataset of aerial action recognition for SAR. In addition, deep-learning models uses these dataset for training in different type of task such as classification and localization. Deep-learning models used for these task can automatically extracts the feature. Out of all other neural networks used for classification or localization, convolution neural networks (CNN) suits more for image-based feature extraction.

* Corresponding author.
E-mail addresses: Balmukund.mishra92@gmail.com (B. Mishra), deepakgarg108@gmail.com (D. Garg), pratik.narang@pilani.bits-pilani.ac.in (P. Narang), vkm.iiti@gmail.com (V. Mishra).

https://doi.org/10.1016/j.comcom.2020.03.012
Received 30 November 2019; Received in revised form 26 February 2020; Accepted 9 March 2020
Available online 14 March 2020

In CNN, each layer uses a convolution filter for feature extraction. An example of two different type of such task is represented in Fig. 2. In this, detection is a combination of classification and localization. The classification problem of images is mainly to classify the image into a different category (labels), while the objective of detection is to identify the label of the object as well as to determine the exact position of classified labels in that image.

As dataset plays a crucial role in the performance of model, this paper proposes a unique dataset of aerial action recognition for SAR. Also, in the aerial surveillance, since, human appears very small and existing algorithms are not able to identify the action performed by them, this paper also proposes a modified action detection model for aerial action detection. The main contribution of the paper is as follows:

- In order to develop any application using deep-learning, the primary requirement is availability of labelled dataset. But for automated search of human using drone surveillance there is no such dataset available in literature. Therefore, in this paper, we have proposed a novel dataset to search humans in rescue for disaster management application.
- Proposed dataset is annotated for two different set of action and is available in the form two action dataset and six action dataset for SAR.
- In addition, an experimental analysis of deep learning object detection models such as Faster R-CNN, R-FCN, and single shot detection (SSD) applied to existing aerial action detection dataset [3] and proposed dataset, has been presented in the paper. Moreover, an modified SSD has also been proposed for better performance in aerial surveillance.

2. Related work

Here, we briefly introduce the current work in the field of dataset and the application of model for aerial human action recognition.

2.1. Aerial action dataset

Several dataset has been developed to solve the problem of action recognition [2,4] from ground view, which is a different scenario compared to the aerial view. A complete list of recently published dataset and their descriptions are available in [5,6]. For aerial action recognition, few datasets have been proposed in the past [3] with 13 actions of hand and other different body parts. To achieve application-level performance, the dataset is required to be precise for action recognition. Existing dataset for aerial action recognition has no action which can help in SAR. Also, the images were captured from a fixed height camera from where the human features not clear for action classification. Detecting actions precisely from that height is almost unrealistic, which is represented by the result in the literature. Some other related dataset is proposed in [3,7]. In addition, another dataset is proposed in [8] for 13 different hand signals used to handle aircraft. This dataset has 13 different actions, having more than 5000 image frames annotated in video format. This dataset has action related to a different pose of hands, such as wave-of, move left. This dataset has specific human actions required for aircraft at the time take-off and landing. Controlling the aircraft landing and take-off is a different scenario compared to drone surveillance for SAR. Instead of single human performing action in a frame, multiple humans performing different action is required for real-life applications.

2.2. CNN for aerial action recognition and object detection

Among the limited work available in the literature to solve the problem of aerial object detection, we have represented the summary of previous attempts made for various object detection and action recognition in Table 1. The need for performing accurate and real-time human detection and action recognition in aerial surveillance has sparked significant research in the past few years. Due to the availability of multiple types of aerial vehicle such as drone, aircraft, air balloons have opened the door for any surveillance in a remote area also. The basic object detection literature can be divided into two categories: Two-stage detection and one-step detection. In the former category, input images in conjunction with thousands of object proposals created by selective search methods are given to a neural network. This network extracts the features from the object proposals, and finally, a classifier determines if there is an object in the proposal or image patches. Region-based methods include RCNN [9], SPP-Net [10], Fast R-CNN [11], Faster R-CNN [12], and Mask R-CNN, and RFCN [13]. Another category is the one-step approach, which is also called regression and classification-based OD, which mainly includes YOLO [14] and SSD [15]. Among the deep learning-based object detectors, SSD has outperformed, because it detects objects in a multi-scale framework. The first part of SSD is made of VGG16 network, and then some extra layers are added in the network in a hierarchical way where first layers contribute more in detecting smaller objects, while the final layers focus on larger objects.

These deep object detection models have been applied for various aerial surveillance application. A detailed survey of detection methods applied for aerial images is provided by [16]. Vehicle detection is an essential application of transportation for traffic monitoring and other related tasks and is discovered by various researchers in the past. VEDAI512 [17] and DLR 3K [18] are some standard datasets for vehicle detection from aerial view. R-CNN, Faster R-CNN, and SSD, three different object detection methods are tested in [19] and found that R-CNN with their proposed region proposal network outperforms. Another relevant dataset is Munich [20] for vehicle detection in aerial imagery, which has been used in [21]. However, this dataset has only 20 images for training. This paper gives a separate region proposal network called AVPN (accurate vehicle proposal network) to hypothesize vehicle locations. Other than vehicles, researchers have tried other objects also for detection, f-ex in [22] YOLO based model have been tested for their dataset created for aeroplane detection. This model uses a five-stage framework that is window evaluation, extraction, and encoding of features, classification, and post-processing as well as extraction of the region of interest. In some work, techniques were applied for detecting objects like birds, drones, and trees. However, human detection and their action recognition from the aerial image is a much harder problem because of its ground covering area. In literature, limited work is available for human detection and their action recognition on aerial images. In [3], a dataset is proposed with 12 actions, and it followed the SSD based approach for action classification. However, more efforts have been made with the same SSD model in [7], and they are getting the performance of 0.28 mAP. Still, the area of action detection in aerial drone images is new and open, needs more work in terms of data as well as models that can detect the humans more accurately and classify their action precisely to be implemented in real-life drone applications.

Other than this aerial dataset and their applied methods, there are other application area where this object detection methods are useful. One of those application areas of object detection is obstacle detection for visually impaired people. In one of the works for visually impaired people [23], a computationally fast and straightforward obstacle detection technique was proposed for real-time obstacle detection using a vision-based sensor. In this, a ground plane removal method was also proposed to filter out the ground area from the image. In another work for obstacle detection, a kinetic sensor and 3D image processing-based system is proposed for the indoor environment [24]. This 5-step process is used, beginning from data acquisition to real-time different type of obstacle detection such as (door, floor, staircase, and wall). There are some other works that we can relate to this research, such as in [25], and a solution is proposed for cyber-threat detection in smart vehicles and its communication using UAV edge computing. In this, a data-driven transportation optimization model is proposed in which cyber-threat detection in smart vehicles is done using a probabilistic

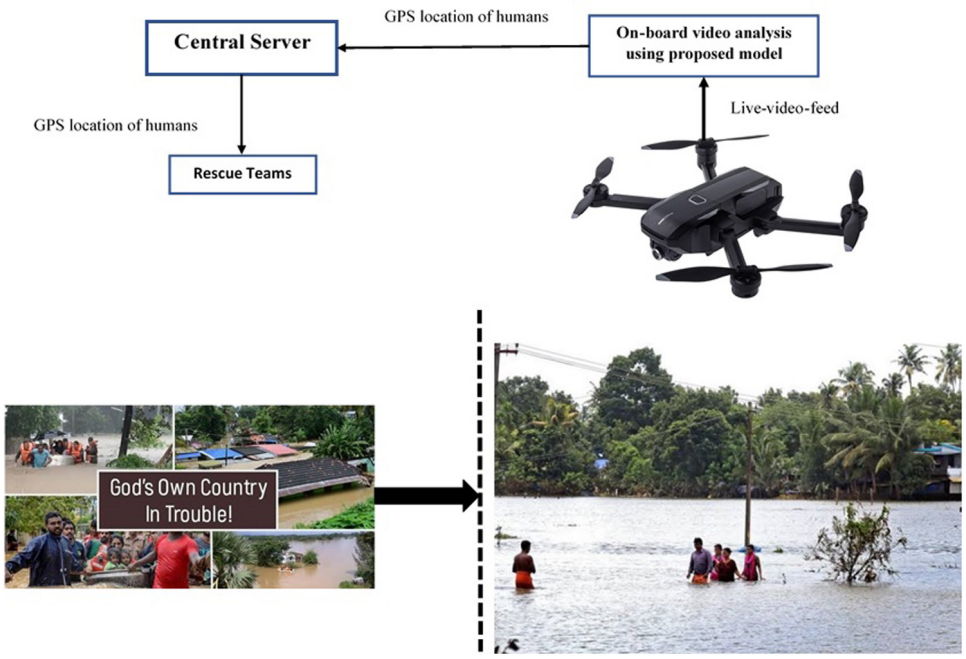

Fig. 1. Example of autonomous search operations for SAR using drone surveillance.

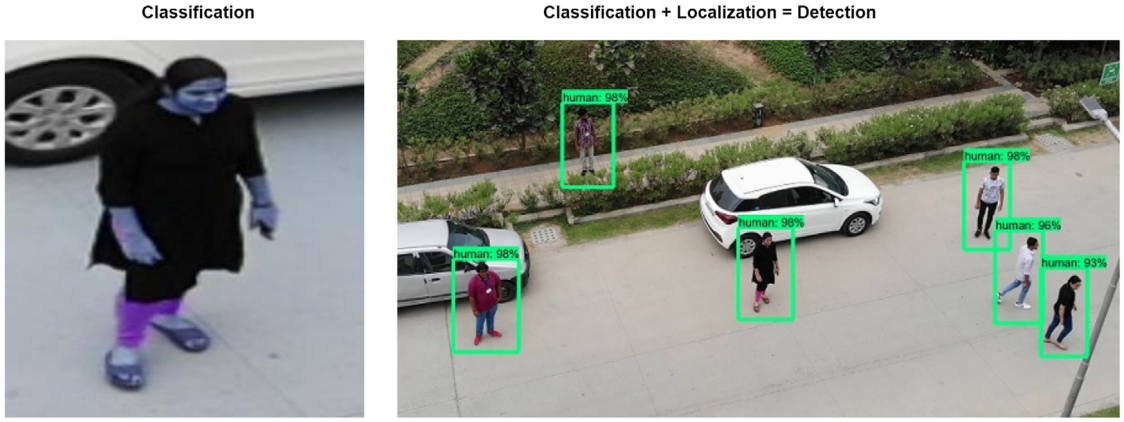

Fig. 2. Visual example of classification and detection [a. represents the outcome of a classification model where model is used to inference the images as human, cat, dog (class-label), while, b. represents the example of detection which is a combination of classification and localization. Bounding boxes around the instance of human represents the presence of human with classified labels around the bounding box.].

data structure. For autonomous drone surveillance, efficient path planning is another area, and for this, in [26], the author has proposed an MVO (Multi-verse optimizer) based optimization technique. This MVO based algorithm performed better in both the form for better fitness value as well as for average execution time. In another research [27], related to drone and its communication, since UAV networks are vulnerable to multiple types of attack, a tree-based attack defence model is proposed for risk assessment. This is a game-theoretic type of solution for assessment, which will keep improving over time. This paper uses a DOS attack to analyse the proposed model using a case study.

3. Dataset development

This section describes the dataset collection, type of action recorded, pre-processing, and the usefulness of dataset for vision-related real-life applications.

3.1. Dataset collection

The data was collected in various locations (in and out of our campus). For this, we have used a drone equipped with a high definition camera from the height between 10 m to 40 m. For video recording, we have used a GoPro Hero 4 Black camera with HD lens (5.4 mm, 15 MP, IR CUT) and a 3-axis solo gimbal. We provide the videos with HD (1920*1080 pixel) formats at 60 fps. The participants were asked to act individually so that in a single image frame, multiple actions or multiple instances of actions can be recorded. A total of 6 actions have been recorded while the drone is moving in the horizontal and vertical direction both. Proposed dataset consists of all the scenarios of human acting, i.e., the human is alone and acting, group of the human performing the same action, and a group of humans performing different actions individually. While recording the video, sometimes wind affected the drone, and different height of video capturing for multiple humans performing different actions makes it closer to the practical scenario. Fig. 3 represents the sample of images used in proposed dataset.

Table 1
Comparison table for object detection and action recognition methods in drone images.

S. No	Title	Type	Methodology	Dataset	Remarks
1	Comprehensive analysis of deep learning based vehicle detection in aerial images [19]	Aerial image vehicle detection	SSD, Fast RCNN faster RCNN	VEDAI512 and DLR 3K dataset	88.7% precision, Computationally expensive and bigger object is easy to detect compared to human.
2	Performance comparison of deep learning techniques for recognizing birds in aerial images [28]	Aerial image Small object Object detection	YOLO, SSH Tiny face	LBAI dataset	Dataset is not sufficient for generalization. Model was trained on low resolution images.
3	Towards fast and accurate vehicle detection in aerial images using coupled region-based convolution neural network [21]	Aerial images Vehicle detection	AVF, SS+RCNN, Faster R-CNN	Munich dataset	Proposed model is failed for hard cases i.e. hard to apply for aerial surveillance
4	Using deep learning and low-cost RGB and thermal cameras to detect pedestrians in aerial images captured by multirotor UAV [22]	Aerial Images Human detection Pattern recognition	HOG+ SVM CNN	GMVRT-v1, GMVRT-v2, and UCF-ARG dataset	90% accuracy, Not suited for multi-person in a single frame UAV surveillance
5	Object recognition in aerial using convolution neural network [29]	Aerial images Aeroplane detection	YOLO	Publicly not available	84% Precision for aeroplane
6	Car detection in aerial images of dense urban areas [30]	Aerial images Vehicle detection Traditional approach	Sliding window approach, with CNN	Vaihingen dataset with Geo-information	0.75 precision Hard-cases are with complex background Proposed model is computationally expensive
7	Region proposal approach for human detection on aerial imagery [31]	Aerial images Pattern recognition Human detection	CNN model for ROI	Publicly not available	As dataset is not available publicly, proposed model need to be checked with publicly available dataset.
8	Okutama-Action: An aerial view video dataset for concurrent human action detection [3]	Aerial images Human detection Action recognition,	SSD	Okutama dataset is proposed	0.18 mAP@0.50IOU
9	Convolution neural networks for aerial multi-label pedestrian detection [7]	Aerial images, Action detection	SSD 500	Okutama	0.28 mAP@0.50IOU
10	UAV-Gesture: A dataset for control and gesture recognition [8]	Aerial images, Action detection	P-CNN descriptor	Dataset is proposed with 13 actions of hand signals	85% accuracy

Fig. 3. Humans performing different action [Sub-figures d, g, h, and i] include the action waving hand which is a symbol of help situation. Sub-figures includes sample of images used to develop our 6-class action dataset captured using drone.

3.2. Action selection

The actions were selected from a general crowd behaviour for UAV capturing the signals. Selected six actions were shown in Fig. 3. Our primary concern is to collect the drone dataset that is helpful for SAR applications. Actions were selected based on the existing action classification dataset by adding one more action as waving a hand. Waving hand action itself has variation such as single hand waving,

Table 2
Summary of proposed six-class action dataset.

Feature	Values
Number of actions	6
Number of images	7000
Average instance per class	1000
Frame rate	60 fps
Resolution	1920*1080
Camera motion	Yes, slow and steady
Annotation and its format	Yes, Bounding box, .xml format

both hand waving, person waving hand while standing at a fixed place, and person waving both hands while he is standing. The same four cases with the person were walking and waving, sitting and waving, and running with waving. So, proposed dataset has a rich amount of sample and variation for our prime class that is waving a hand. For the action selection, some more factors have been taken care as follows:

- They should be easily identifiable from a moving drone camera.
- Actions need to be crisp enough to differentiate from each other.
- Actions should represent the normal crowd behaviour.
- Actions should be diverse enough to meet the requirement for different real-life applications.
- The action should be discriminative, so that spatial and temporal both methods can be applied.

3.3. Variation in dataset

Actors who participated in the dataset are not professionals for drone-based signals and performing actions in-crowd. Since we have captured the data with different actors, performing a different action in all the videos, so there is a variation in the way different people performing the same action. Also, the dataset is captured from the different orientation, angle, camera movement, and height make it more generalize for real-time applications. Videos are captured at different time of day, and different lighting that gives a variation in terms of lightening condition to the dataset. Our actions are selected for multiple real-life applications that require some challenging actions which are difficult to classify, like running and walking. The dataset is rich in variety, since, it is captured at different days of month with actors having different clothing styles, and style of acting. These variations create a challenging dataset for action recognition; at the same time, proposed dataset is diverse enough to be close as the real-time practical scenarios.

3.4. Pre-processing of dataset

Dataset for drone-based action recognition is collected in the form of video. Since we are targeting for only spatial feature-based multi-class action recognition, all the frames are not important. To avoid repetition, we skipped every 10 frames for one frame in proposed dataset. Other than our proposed dataset, existing Okutama dataset is also re-framed according to our requirement. Okutama dataset is originally published for 13 action. For our experimental study, we have prepossessed the Okutama videos by extracting the frames and annotating them. The actions we have annotated is same as it was originally in Okutama, one action that is waving hand was added extra in the annotation.

Preprocessing steps for dataset development after capturing video is as follows:

- **Frame extraction**: Dataset was captured in the form of video and converted in the form of images bu extracting its frames using the OpenCV package of python.
- **Frame selection**: Originally, the extracted frames have repetitions in the form of action features. We have checked by removing 10, 15, 20, frames for keeping one frame in proposed dataset. In the end, we have decided to keep one frame after every 10 frames after experimenting at different levels and our previous experience.
- **Action annotation**: For deep learning models to work accurately, each human must be localized in the image. Fig. 4 shows the sample of annotations. In this, there are a varying number of humans acting differently in different frames of video. However, for other applications of UAV, where a simple human detector is required, this dataset can be utilized with a minor change in.CSV file of the dataset. For the annotation of every image, labelImg application is used. Finally, we have two sets of annotations in which two and six actions were annotated for SAR UAV dataset. Fig. 4 represents the instances of humans annotated as different actions.

  For our action detection dataset, the video is captured from different height through our drone. For two class-action detection dataset, frames are extracted from the video by skipping every 12 frames. Our data contains approx. 500 samples of each class of action. The labelImg application annotates each frame. Our six class-action detection data-et also contains more than 200 samples and annotated by the same labelImg application.

3.5. Dataset summary

There are many datasets available for aerial object detection and action recognition in literature. For aerial image and video dataset, Table 3 represents a comparison of the state-of-the-art dataset. Best-suited data for our drone-based action recognition is Okutama action dataset [3]. For our analysis, we have modified the Okutama dataset into images by extracting their frames and annotated them as per our requirement.

Also, we have developed the dataset, as explained above. For human action detection dataset, we are using our drones to capture images and videos, and, in this process, we have captured two different kinds of images with two actions and six actions. Existing dataset for human and their action detection is very complicated and take from more than 65-metre height, and results are not satisfactory with the dataset, as shown in our experiment as well as in state of the art. So, for SAR and other surveillance applications using a drone camera, this dataset is not useful. The six actions of our second part of the action detection dataset are:

- Person Standing
- Person Sitting
- Person Laying
- Person Handshaking
- Person Walking
- Person Waving

The summary of this six-class action dataset is given in Table 2. For a two-class action dataset, we must combine other actions in a single class other than waving a hand. So, our two-class action dataset contains the following classes:

- Person waving
- Other

Initially, dataset is captured in the form of a video using the drone. As drones are moving, compared to others, in proposed dataset human features are being explored in a better way from the height 10 to 40 m. At this height, features are more visible and can contribute to the action detection and classification applications well.

4. Performance evaluation metrics

mAP and IOU are the standard COCO evaluation parameters used to the evaluate object detection models. Hence, to compare our proposed model with the state-of-the-art models in literature, these parameters are suitable. The details of these parameters are discussed in this

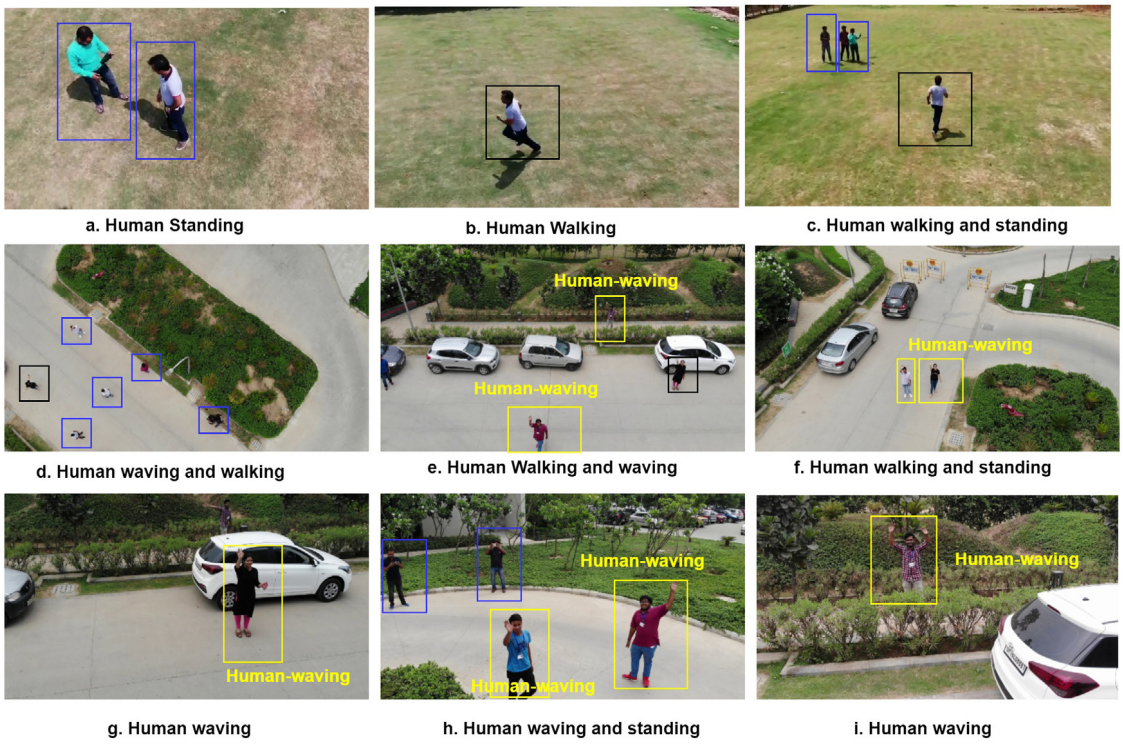

Fig. 4. Annotations of humans performing multiple actions [yellow colour bounding box represents the waving hand]. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Comparison of proposed dataset with the state-of-the-art datasets.

Dataset	Scenario	Purpose	Environment	Frames	Classes	Resolution	Year
UT Interactions [32]	Surveillance	Action-recognition	Outdoor	36K	6	320*240	2010
NATPOS [33]	Aircraft-signalling	Gesture-recognition	Indoor	N/A	24	320*240	2011
VIRAT [34]	Drone-Surveillance	Event-recognition	Outdoor	Many	23	Varying	2011
UCF101 [2]	YouTube	Action-recognition	Varying	558K	24	320*240	2012
J-HMDB [35]	Movies, YouTube	Action-recognition	Varying	32K	21	320*240	2013
Mini-Drone [36]	Drone	Privacy-protection	outdoor	23.3K	3	1920*1080	2015
Campus [37]	Surveillance	Object-tracking	outdoor	11.2K	1	1414*2019	2016
Okutama-action [3]	Drone	Action-recognition	outdoor	70K	13	3840*2160	2017
UAV-gesture [8]	Drone	Gesture-recognition	outdoor	37.2K	13	1920*1080	2018
Proposed-dataset	Drone	Action-recognition	outdoor	30K Human instance sorted out of 75K frames	6	1920*1080	——

section. Also, as precision and recall are two basic parameters on which all these evaluation parameters depends, a brief description of these parameters are also given in this section.

4.1. Intersection over union (IOU)

Intersection over Union is an evaluation metric used to measure the accuracy of an object detector on a dataset. We often see this evaluation metric used in object detection challenges, such as the popular PASCAL VOC challenge. Intersection over Union is simply an evaluation metric. Any algorithm that provides predicted bounding boxes as output can be evaluated using IoU [38]. More formally, to apply Intersection over Union to evaluate an (arbitrary) object detector we need:

- The ground-truth bounding boxes (i.e., the hand-labelled bounding boxes from the testing set that specify wherein the image our object is).
- The predicted bounding boxes from our model.

Fig. 5 represents the calculation of IOU value according to the ground truth and detected bounding boxes.

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \tag{1}$$

4.2. Mean average precision (mAP)

The mAP is the metric to measure the accuracy of object detectors like Faster R-CNN, SSD, etc. It is the average of all the average precision

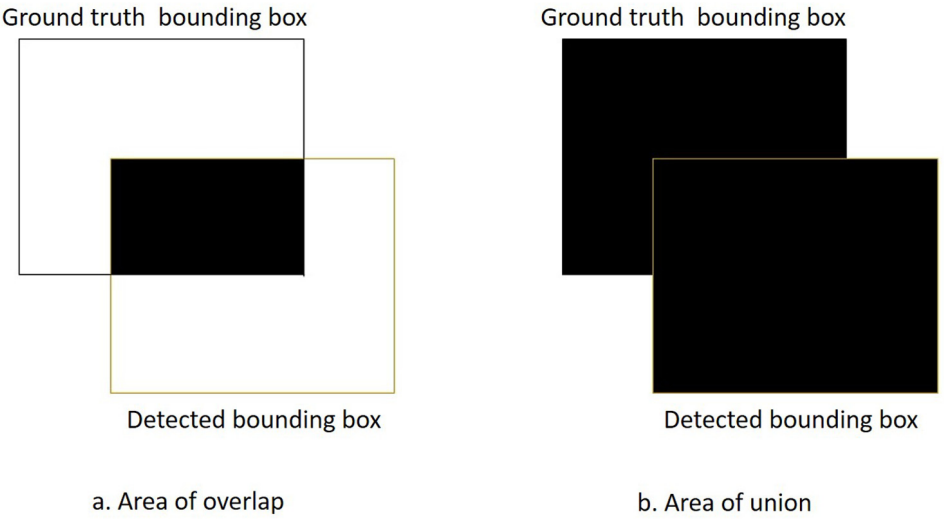

Fig. 5. Visual example of intersection over union.

(AP) calculated for all the classes [39].

$$mAP = \frac{1}{n}\sum_{i=0}^{n} AP(i) \tag{2}$$

n represents the number of classes in the dataset.

Average precision is then calculated by taking the area under the precision–recall curve. This is done by segmenting the recalls evenly into 11 parts for different IOU values between 0 to 1.

$$AP = \frac{1}{11}\sum_{0.0}^{1} Pr \tag{3}$$

4.3. Precision and recall

Precision measures how accurate your predictions are. i.e., the percentage of your positive predictions are correct.

Recall measures how good you find all the positives. For example, we can find 80 percent of the possible positive cases in our top K predictions.

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

$$Recall = \frac{TP}{(TP + FN)} \tag{5}$$

TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative

5. Proposed framework

Proposed architecture with the developed dataset for action detection in drone surveillance, can be used to identify the situations where humans are asking for help. As shown in Fig. 1, on board autonomous analysis of drone images can quickly find humans stuck in the disaster prone area. Proposed dataset is generalized and have enough variation to be used for the automation of such application. In addition, the architecture of proposed model for detecting the action is shown in Fig. 6. Proposed model use the feature of multiple level of convolution for the classification of object. Specially, the feature of initial convolution layer contribute more for the classification of small object. In case of drone surveillance images taken from the top, object appears very small and their features are less explored for classification. So, proposed model utilize the feature of initial convolution layer, that improves the result of proposed model for action recognition in drone surveillance.

The block-diagram for SAR shown in Fig. 1 is based on action detection of humans stuck in disaster-affected area and for this a unique model is proposed. The proposed model for action detection is inspired by the pyramidal feature extraction [15] and utilization for classification and localization. Fig. 6 shows the architecture of proposed model for multi-class action detection based on spatial features. As shown in Fig. 6, proposed model uses the feature of different convolution layer for the localization task, specially the features taken after 3rd convolution layer is useful for smaller objects. So, we have experimented with the various combination of convolution network such as VGG16, Inception and the feature of various layer for the localization. After the analysis of these experiments, we found, the Inception network with the architecture shown in Fig. 6 is performing better. In the proposed architecture of action detection model, feature of the 3rd convolution layer is fed directly to the detection generation which is a key factor of the performance improvement. Also four extra convolution layer is used after inception network where the features are fed to the detection generation in pyramidal way. The parameter value of these extra convolution network is same as the convolution layer in Inception network. Details of hyper-parameters of proposed model is given in Table 4. as our proposed model is trained for two class and six class (both) dataset, with activation function as relu. Other parameters are used after optimization as given in Table 4 and details of results are given in Section 7.

Table 4
Details of hyper-parameters and their values in final trained model (Proposed model).

Hyper-parameters	Values
Number of classes	2 and 6
Activation	relu
Batch-normalization	Yes
IOU	0.5
Batch size	24
Optimizer	rmsprop
Momentum-optimizer value	0.9
Initial learning rate	0.004

In our proposed approach, the drone will scan the disaster-affected area, and at the same time, our proposed model deployed on the drone will recognize the specific action as a help situation. Our proposed approach for SAR follows the intuition of human nature as they wave a hand in the direction of the aerial vehicle as a symbol of asking help. This specific gesture of waving hand has a variety of action such as human-standing and waving hand, human-laying and waving hand, human walking and running hand, human running, and waving hand. In other words, for human detection and action recognition in a given drone image can be precisely done by identifying the feature at each

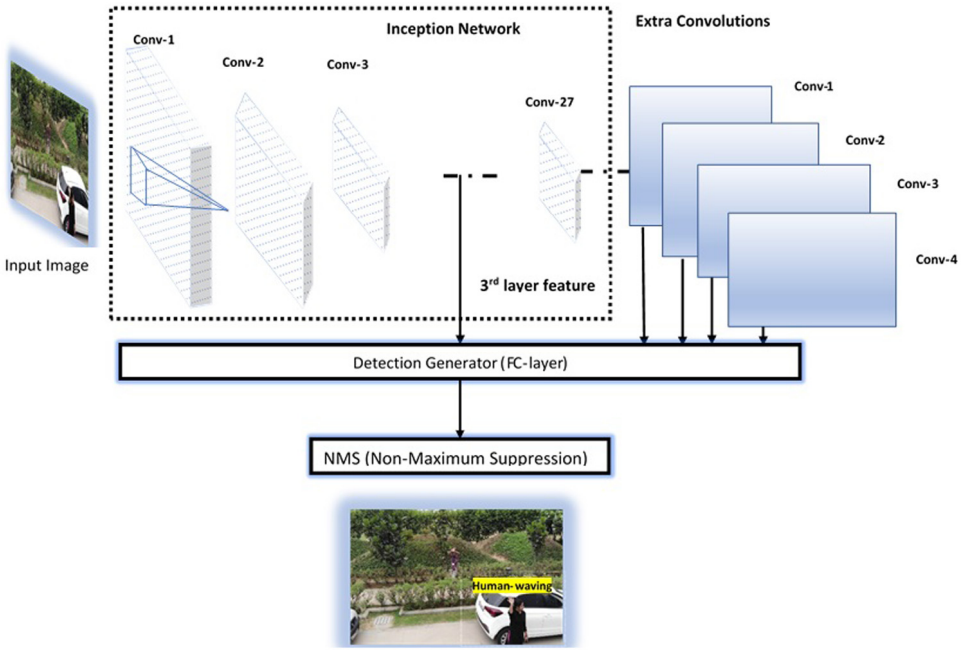

Fig. 6. Architecture of the proposed action detection model [In detection network, FC layer represents the dense layer of the network, which is used with softmax function for final feature classification. In the end, NMS is a function in computer vision that uses a threshold value and gives an output of single bounding box with input as multiple bounding boxes on the object.].

layer of convolution. Its feature map is expected to consist of multiple structures and features of body parts and their pose, which corresponds to the human action. The proposed implicit model exploits the feature of every CNN network in a hierarchical pyramidal way. At the end of our trainable network, NMS (Non-maximum suppression) is used to select the best bounding box based on its threshold score. Our proposed model is applied to the detection of two-class and six-class action. It is also applied for the publicly available Okutama dataset [3], and result is discussed in the Section 7.

6. Experimental setup

Experiments were performed on the NVIDIA DGX-1 V100 supercomputer having 7.8 TFLOP/s for FP64 computation power. To recognize human action, state-of-art object detection was applied to the proposed dataset. In addition to this, experiments were performed on publicly available Okutama dataset as well. Both datasets have frames that contain multiple people performing different actions simultaneously. As the image size of both datasets is equal, i.e., 1920 * 1080 pixels, the results of the various model will give a good comparison between the actions performed in both the dataset. The results of these models are compared based on the standard Coco performance metrics (mAP and IOU). Our proposed dataset for two-class and six-class actions contains 6000 images. For two-class experiments, actions are hands-up and others. Second class 'other' include all the annotated class except hands-up. Dataset was split into training and testing in the following ways. For training and testing, out of 6000 images, 2000 images were selected randomly with their annotation in XML format. Out of 2000 images, 1800 images are used for training, and the other 200 are used for testing. For evaluation of the trained model, for every step of evaluation, randomly 10 images were selected by using random shuffle. All three models were trained and tested on both the dataset with this configuration of the dataset split in the TensorFlow environment.

Our proposed model is inspired by the SSD of object detection because its performance in terms of mAP is comparatively equivalent in relatively very less inference time. Our proposed model is tuned for different hyper-parameters such as batch size, number of steps, learning rate, and optimizer. Starting from our first model (Faster R-CNN), we are using the ResNet classification model, with the first stage feature size as 16. As for our requirement of processing unit, we have varied the batch size from 32 to 8 and the learning rate from 0.003 to 0.002. Initially, the model will be trained for learning rate 0.003, and after 11 000 steps, the learning rate will automatically decrease to 0.002.

7. Results and analysis

In this section, we have discussed the results in detail based on the availability of visual and statistical results for human detection and action recognition.

Table 5 shows the performance of deep learning object detection models applied on publicly available Okutama dataset. The performance is evaluated on a standard coco evaluation metric (mAP). Our result shows that faster R-CNN is performing comparatively better on this dataset. In addition to this, Table 6 shows the results of models applied to our proposed dataset. Tables 5 and 6 are both shows the result with six-class. A six-class dataset is developed for general surveillance applications. For our next set experiments, which are intended for SAR, the proposed dataset is converted into two class datasets. Table 7 shows the result of deep learning models applied in a two-class action dataset, which is helpful for SAR applications. For the analysis of our results shown in Tables 5, 6, and 7, we have analysed the result qualitative analysis of the visual inspection of the bounding boxes. Also, the qualitative analysis of the result is done based on the evaluation parameters mentioned here.

7.1. Qualitative analysis

Visual results of the experiments of applying deep learning models on the proposed dataset are shown in Fig. 7. In addition to this, models were applied to publicly available Okutama dataset for 6-action. Fig. 3 Shows the distribution of classes and samples available in proposed dataset. It shows that in proposed dataset sufficient samples are available for models to learn the feature accurately for classification. Fig. 3 also shows the visuals of proposed dataset, where each action is shown; however, due to the presence of multiple actions in the same scene, the sub-images have captions that include multiple actions. In addition

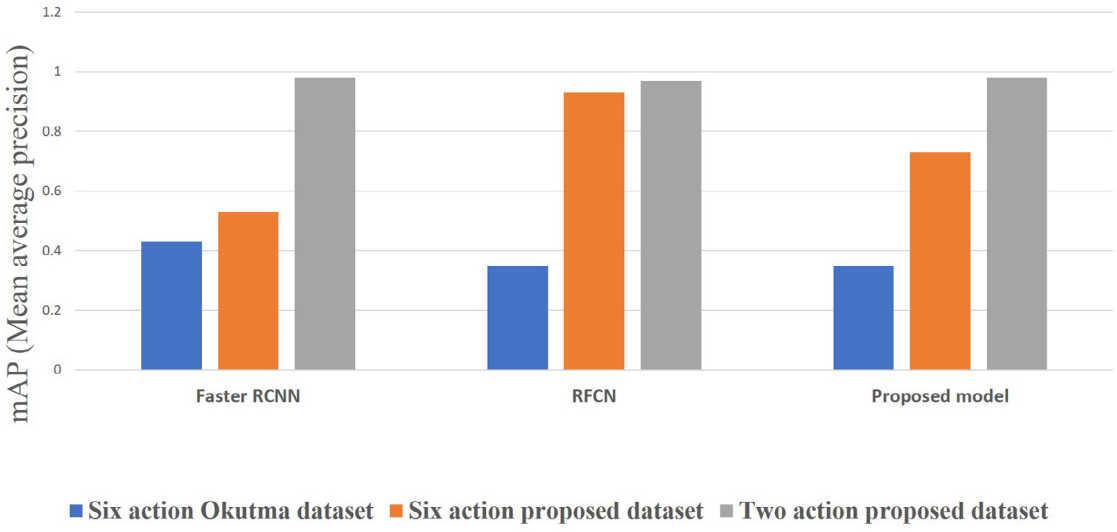

Fig. 7. Comparison of deep learning models for action recognition in aerial images.

to this, Fig. 4 shows the annotation of each action inside the image frame. In this figure, sub-images include captions, and inside that sub-image, each bounding boxes of the different classes have a different colour. Among all the classes, our main objective is to identify the waving hand, which is shown by the yellow colour bounding box. Other implementation results for models applied to this dataset are discussed and analysed in the next section.

7.2. Quantitative analysis

Quantitative evaluation methods for action recognition depends on the approach we have followed. As we have applied the model for multi-human action recognition, the standard evaluation metrics are average precision and mean average precision with different IOU values. COCO standard performance measuring parameter is mAP (mean average precision), which is the average of all the best recalls of all the classes. In this evaluation method, the detected bounding box is considered valid if it matches a corresponding ground truth bounding box, and there should not be any other bounding box with the same ground truth. The second metric that is used to evaluate the performance of modes average precision (AP).

Table 5 compares the performance of deep learning models applied to six class actions of Okutama dataset. From these results, comparatively, SSD is performing better, as the mAP value is comparatively equivalent, while the inference time of SSD is very less than the other two models. Table 6 shows the result of deep learning models applied to six class actions of our proposed dataset. It shows the validity of our proposed dataset for the real-time application. The mAP value for all three models is higher than 90 percent. It can also be concluded from this, that SSD is favourable, and proposed dataset is captured from a good enough height from where features are visible to be identified. Table 7 represents the experimental result of the selected models applied to our two-class action dataset. As it is mentioned in the dataset development section also, this experiment is performed for SAR applications, where we have to identify the waving hand. The experimental result shows the validity of our approach that waving had can be identified using these object detection models. Our trained model can be used for classifying the action as waving and non-waving. Among all those models that we have used here, SSD looks better as it gives comparatively equivalent with less inference time. We have also presented the result using the comparison graph represented in Fig. 7. The graph shows the result of model performance applied in all three datasets, including publicly available Okutama dataset. It shows that the models are performing better on our propose dataset as the selected actions were different enough to be discriminated by the deep learning models. Also, the height from where images were captured is another factor of this performance.

Table 5
Performance of object detection models for action detection for six-classes of Okutama dataset.

Models	22 000 Steps mAP @ 0.50 IOU	200 000 Steps mAP @ 0.50 IOU	500 000 Steps mAP @0.50 IOU
Faster R-CNN	0.43	0.38	0.35
R-FCN	0.35	0.32	0.20
Proposed model	0.20	0.35	0.32

Table 6
Performance of deep learning models for action detection on proposed six-class aerial action dataset.

Models	mAP @ 0.50 IOU
R-FCN	0.93
Faster R-CNN	0.53
Proposed model	0.73

Table 7
Performance of deep learning models for action detection on proposed two-class aerial action dataset.

Models	mAP @ 0.50 IOU
Faster R-CNN	0.988
R-FCN	0.97
Proposed model	0.98

8. Conclusion

In this paper, we have proposed a drone dataset for human action recognition. This dataset can also be used for human detection and other such task for different surveillance applications. Proposed dataset has a rich amount of variety in terms of colour, height, actor, and background. This variation makes it generalized for proposed dataset to be used for various applications. In addition, as our primary objective is to provide the support for SAR using drone surveillance, we have presented an experimental comparison of deep-learning action detection model applied on proposed dataset and other publicly available dataset. This paper also proposes a novel detection model for action recognition, and it achieves 7% higher mAP value in comparison with the state-of-the-art SSD model when applied to publicly available Okutama dataset.

Also, when the proposed model is applied on our two-class action detection dataset for SAR, it achieves 0.98 mAP value, which is a decent performance value for real-time application. Proposed dataset is available on the link https://www.leadingindia.ai/data-set.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Balmukund Mishra: Conceptualization, Methodology, Writing - original draft. **Deepak Garg:** Data curation, Investigation, Supervision, Writing - original draft, Project administration. **Pratik Narang:** Visualization, Investigation, Writing - original draft. **Vipul Mishra:** Data curation, Formal analysis, Supervision, Writing - original draft, Visualization.

References

[1] Zacc Dukowitz, Drones in search and rescue: 5 stories showcasing ways search and rescue uses drones to save lives, 2019, https://uavcoach.com/search-and-rescue-drones/. (Accessed 3 Dec 2019).

[2] Khurram Soomro, Amir Roshan Zamir, Mubarak Shah, UCF101: A dataset of 101 human actions classes from videos in the wild, 2012, arXiv preprint arXiv:1212.0402.

[3] Mohammadamin Barekatain, Miquel Martí, Hsueh-Fu Shih, Samuel Murray, Kotaro Nakayama, Yutaka Matsuo, Helmut Prendinger, Okutama-action: An aerial view video dataset for concurrent human action detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 28–35.

[4] Joao Carreira, Andrew Zisserman, Quo vadis, action recognition, a new model and the kinetics dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6299–6308.

[5] Soo Min Kang, Richard P. Wildes, Review of action recognition and detection methods, 2016, arXiv preprint arXiv:1610.06906.

[6] Pramod Kumar Pisharady, Martin Saerbeck, Recent methods and databases in vision-based hand gesture recognition: a review, Comput. Vis. Image Underst. 141 (2015) 152–165.

[7] Amir Soleimani, Nasser M. Nasrabadi, Convolutional neural networks for aerial multi-label pedestrian detection, in: 2018 21st International Conference on Information Fusion, FUSION, IEEE, 2018, pp. 1005–1010.

[8] Asanka G. Perera, Yee Wei Law, Javaan Chahl, UAV-GESTURE: a dataset for UAV control and gesture recognition, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018.

[9] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, Arnold WM Smeulders, Selective search for object recognition, Int. J. Comput. Vis. 104 (2) (2013) 154–171.

[10] Pulak Purkait, Cheng Zhao, Christopher Zach, SPP-NEt: Deep absolute pose regression with synthetic views, 2017, arXiv preprint arXiv:1712.03452.

[11] Ross Girshick, Fast r-cnn, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.

[12] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99.

[13] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, Xiang Ruan, Saliency detection with recurrent fully convolutional networks, in: European Conference on Computer Vision, Springer, 2016, pp. 825–841.

[14] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.

[15] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C Berg, Ssd: Single shot multibox detector, in: European Conference on Computer Vision, Springer, 2016, pp. 21–37.

[16] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, Deva Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2009) 1627–1645.

[17] Lars Wilko Sommer, Tobias Schuchert, Jürgen Beyerer, Fast deep vehicle detection in aerial images, in: 2017 IEEE Winter Conference on Applications of Computer Vision, WACV, IEEE, 2017, pp. 311–319.

[18] Tao Qu, Quanyuan Zhang, Shilei Sun, Vehicle detection from high-resolution aerial images using spatial pyramid pooling-based deep convolutional neural networks, Multimedia Tools Appl. 76 (20) (2017) 21651–21663.

[19] Lars Sommer, Tobias Schuchert, Jürgen Beyerer, Comprehensive analysis of deep learning based vehicle detection in aerial images, IEEE Trans. Circuits Syst. Video Technol. (2018).

[20] Kang Liu, Gellert Mattyus, Fast multiclass vehicle detection on aerial images, IEEE Geosci. Remote Sens. Lett. 12 (9) (2015) 1938–1942.

[21] Zhipeng Deng, Hao Sun, Shilin Zhou, Juanping Zhao, Huanxin Zou, Toward fast and accurate vehicle detection in aerial images using coupled region-based convolutional neural networks, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 10 (8) (2017) 3652–3664.

[22] Diulhio de Oliveira, Marco Wehrmeister, Using deep learning and low-cost RGB and thermal cameras to detect pedestrians in aerial images captured by multirotor UAV, Sensors 18 (7) (2018) 2244.

[23] A. Jindal, N. Aggarwal, S. Gupta, An obstacle detection method for visually impaired persons by ground plane removal using speeded-up robust features and gray level co-occurrence matrix, Pattern Recognit. Image Anal. 28 (2) (2018) 288–300.

[24] Huy-Hieu Pham, Thi-Lan Le, Nicolas Vuillerme, Real-time obstacle detection system in indoor environment for the visually impaired using microsoft kinect sensor, J. Sens. 2016 (2016).

[25] Sahil Garg, Amritpal Singh, Shalini Batra, Neeraj Kumar, Laurence T Yang, UAV-empowered edge computing environment for cyber-threat detection in smart vehicles, IEEE Netw. 32 (3) (2018) 42–51.

[26] Puneet Kumar, Sahil Garg, Amritpal Singh, Shalini Batra, Neeraj Kumar, Ilsun You, MVO-based 2-D path planning scheme for providing quality of service in UAV environment, IEEE Internet Things J. 5 (3) (2018) 1698–1707.

[27] Sahil Garg, Gagangeet Singh Aujla, Neeraj Kumar, Shalini Batra, Tree-based attack–defense model for risk assessment in multi-UAV networks, IEEE Consum. Electron. Mag. 8 (6) (2019) 35–41.

[28] Yang Liu, Peng Sun, Max R Highsmith, Nickolas M Wergeles, Joel Sartwell, Andy Raedeke, Mary Mitchell, Heath Hagy, Andrew D Gilbert, Brian Lubinski, et al., Performance comparison of deep learning techniques for recognizing birds in aerial images, in: 2018 IEEE Third International Conference on Data Science in Cyberspace, DSC, IEEE, 2018, pp. 317–324.

[29] Matija Radovic, Offei Adarkwa, Qiaosong Wang, Object recognition in aerial images using convolutional neural networks, J. Imaging 3 (2) (2017) 21.

[30] Mohamed ElMikaty, Tania Stathaki, Car detection in aerial images of dense urban areas, IEEE Trans. Aerosp. Electron. Syst. 54 (1) (2017) 51–63.

[31] Željko Marušić, Dunja Božić-Štulić, Sven Gotovac, Tonćo Marušić, Region proposal approach for human detection on aerial imagery, in: 2018 3rd International Conference on Smart and Sustainable Technologies, SpliTech, IEEE, 2018, pp. 1–6.

[32] Michael S. Ryoo, Jake K. Aggarwal, Spatio-temporal relationship match: video structure comparison for recognition of complex human activities, in: ICCV, Vol. 1, Citeseer, 2009, p. 2.

[33] Yale Song, David Demirdjian, Randall Davis, Tracking body and hands for gesture recognition: Natops aircraft handling signals database, in: Face and Gesture 2011, IEEE, 2011, pp. 500–506.

[34] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, JK Aggarwal, Hyungtae Lee, Larry Davis, et al., A large-scale benchmark dataset for event recognition in surveillance video, in: CVPR 2011, IEEE, 2011, pp. 3153–3160.

[35] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, Michael J Black, Towards understanding action recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3192–3199.

[36] Margherita Bonetto, Pavel Korshunov, Giovanni Ramponi, Touradj Ebrahimi, Privacy in mini-drone based video surveillance, in: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, Vol. 4, FG, IEEE, 2015, pp. 1–6.

[37] Alexandre Robicquet, Amir Sadeghian, Alexandre Alahi, Silvio Savarese, Learning social etiquette: Human trajectory understanding in crowded scenes, in: European Conference on Computer Vision, Springer, 2016, pp. 549–565.

[38] Md Atiqur Rahman, Yang Wang, Optimizing intersection-over-union in deep neural networks for image segmentation, in: International Symposium on Visual Computing, Springer, 2016, pp. 234–244.

[39] Paul Henderson, Vittorio Ferrari, End-to-end training of object class detectors for mean average precision, in: Asian Conference on Computer Vision, Springer, 2016, pp. 198–213.