drones

MDPI

Article

Background Invariant Faster Motion Modeling for Drone Action Recognition

Ketan Kotecha [1,*], Deepak Garg [2], Balmukund Mishra [2], Pratik Narang [3] and Vipul Kumar Mishra [2]

[1] Symbiosis Centre for Applied Artificial Intelligence, Symbiosis International (Deemed University), Pune 412115, India
[2] Department of Computer Science and Engineering, Bennett University, Greater Noida 201310, India; Deepak.garg@bennett.Edu.in (D.G.); Bm7477@bennett.edu.in (B.M.); vipul.mishra@bennett.edu.in (V.K.M.)
[3] Birla Institute of Technology and Science (BITS Pilani), Pilani 333031, India; pratik.narang@pilani.bits-pilani.ac.in
* Correspondence: head@scaai.siu.edu.in

Abstract: Visual data collected from drones has opened a new direction for surveillance applications and has recently attracted considerable attention among computer vision researchers. Due to the availability and increasing use of the drone for both public and private sectors, it is a critical futuristic technology to solve multiple surveillance problems in remote areas. One of the fundamental challenges in recognizing crowd monitoring videos' human action is the precise modeling of an individual's motion feature. Most state-of-the-art methods heavily rely on optical flow for motion modeling and representation, and motion modeling through optical flow is a time-consuming process. This article underlines this issue and provides a novel architecture that eliminates the dependency on optical flow. The proposed architecture uses two sub-modules, FMFM (faster motion feature modeling) and AAR (accurate action recognition), to accurately classify the aerial surveillance action. Another critical issue in aerial surveillance is a deficiency of the dataset. Out of few datasets proposed recently, most of them have multiple humans performing different actions in the same scene, such as a crowd monitoring video, and hence not suitable for directly applying to the training of action recognition models. Given this, we have proposed a novel dataset captured from top view aerial surveillance that has a good variety in terms of actors, daytime, and environment. The proposed architecture has shown the capability to be applied in different terrain as it removes the background before using the action recognition model. The proposed architecture is validated through the experiment with varying investigation levels and achieves a remarkable performance of 0.90 validation accuracy in aerial action recognition.

Keywords: drone surveillance; human detection; action recognition; deep learning; search and rescue

Citation: Kotecha, K.; Garg, D.; Mishra, B.; Narang, P.; Mishra, V.K. Background Invariant Faster Motion Modeling for Drone Action Recognition. *Drones* **2021**, *5*, 87. https://doi.org/10.3390/drones5030087

Received: 25 July 2021
Accepted: 25 August 2021
Published: 31 August 2021

1. Introduction

Video action detection has improved dramatically in the last few years, owing largely to the adoption of deep learning action recognition models [1–4] and video databases [5–7]. In addition, many prominent convolution neural networks [8–11] are available in the literature for the image recognition task. However, these CNN's cannot model the motion feature of individuals effectively from a crowd video. Using these CNN's for aerial action recognition can provide a variety of real-life applications for search and rescue using the dataset proposed in [12,13]. Aerial and drone surveillance can also be used for real-time purposes, such as detecting odd behavior in border areas [14], violence and suspicious activity in crowds [15], urban and rural scene understanding. However, due to the intrinsic complexity of aerial footage, motion modeling in action detection remains a difficult task. The variable nature of human characteristics from different angles and heights is one of the primary challenges in aerial or drone monitoring.

The critical need for a practical approach of action recognition is that it must be applied simultaneously from various heights of live stream video, environment, and with single or multiple humans. Various approaches for multi-human action recognition are summarized in [16], with a detailed discussion on opportunities and challenges. The benefit of using drone surveillance is that a wide remote area can be searched in less time. We have recently seen different surveillance tasks, such as using drones for traffic control and crowd monitoring during COVID 19. These activities demonstrate the potential and future of drone surveillance. Automation of such surveillance applications, on the other hand, is essential for taking it to the next level. This automation necessitates on-device video analysis. Also, the capability of the drone is improving in terms of battery life, storage capacity, and processing capability. There is a pressing need for sophisticated algorithms that can detect and recognize persons and their activities and emotions. Most known action recognition algorithms use a two-stage procedure: they estimate the optical flow using the EPE (expected predicted error), then feed the estimated optical flow into the subsequent action recognition module. However, the recent articles [17,18] indicate that this co-relationship is not vital for the overall modeling of action recognition. Due to the availability of adequate data where object features are visible, the deep learning algorithm's performance with ground-level images has recently been improved a lot. Deep learning techniques, on the other hand, are being tested in aerial drone surveillance. The existing action recognition model's performance is subpar in aerial and drone surveillance due to a lack of accurate data on individual persons' temporal actions.

Large visual variances in characteristics, such as occlusions, position variations, and illumination alterations, provide substantial aerial surveillance issues in practical implementation. Deep learning models use human features to understand the shape, texture, and size in the spatial-temporal domain for action categorization. Such qualities are not immediately apparent in drone-captured videos. Robust temporal modeling of the individual is essential for the full presentation of human action traits. This research provides several levels, including a novel architecture, a quick and accurate temporal motion modeling framework, and an upgraded temporal network for action identification to address these issues and employ them for search and rescue in emergencies. The crucial contribution of this paper is as follows:

- To the best of our knowledge, this is the first time a unique architecture has been offered to address the multi-class challenge of activity recognition in aerial and drone surveillance.
- The proposed architecture uses a unified approach for faster motion featuring modeling (FMFM) and accurate action recognition (AAR) working together for better performance.
- A unique five-class-action dataset for aerial drone surveillance was introduced, with approximately 480 video clips distributed in five different classes.
- It developed and trained a unique video-based action detection model that can correctly classify human action in drone videos.

2. Literature

In recent years, search and rescue have received a lot of attention. It might be more effective if deep learning techniques like activity recognition, object detection, and object classification were used. As a result, some critical state-of-the-art research suitable for the automation of search and rescue applications is explored here. Human activity recognition has been a hot topic in recent years, and numerous algorithms for intelligent surveillance systems have been developed in the literature [19]. However, traditional algorithms for action recognition use manual feature extraction [20–23], and hence performance is not up to the mark comparatively. Also, it requires lots of human effort unnecessarily. The emergence of recent deep learning algorithms for intelligent surveillance [24,25] shows the capability of deep learning algorithms. However, applying these models to many humans present in the same scene, especially in an unstructured environment, is complicated, and the article [16] gives a detailed discussion on the complexity and opportunity of this task. Deep learning models are popular nowadays due to the dataset availability, which is

critical for getting adequate performance in the field of action recognition, [25] outlined the current challenges for real-world applications. Analysis of deep learning models shows the key to the success of deep learning models is the availability of datasets, and this enables the models to learn the features required for classification of human action. Several such datasets are available in the literature, such as [26] introduced a video level action dataset called UCF and article [5] proposes a similar dataset called kinetics. However, each dataset has its limitation, which limits its use in different applications. Table 1 presents a summarized review of the literature for some key papers and datasets in drone-based surveillance recently.

These datasets have been utilized to perform profound learning algorithms, such as T-CNN [27], the temporary sector network [28], and the temporary pyramid network [10]. Furthermore, the modeling of movement and appearance is supported by a variant of a 3D convolution network with a model for deformation attention [4]. Models of spatial-temporal feature extraction for action recognition have had tremendous success in recent years; for example, a composition model based on spatially and temporally distinguished elements is suggested in [29]. In addition, [30] recommends a transferable technique for recognizing action using a two-stream neural convolution network. Another field of action recognition that can be useful for search and rescue is hand gesture recognition, and a two-stage convolution neural network is recently proposed in [31].

Other methods of action recognition use the body key points for classifying action to different classes, such as in [32], in which both 2D and 3D features of body key points are used. In [33], a human action recognition model is proposed based on pose estimation, specifically on pose estimated map, and experimented with the ntu rgb + d [34] and UTD-111MHAD [35] datasets. These models for action recognition performed well with ground-level action recognition. However, these algorithms do not perform well in aerial or drone surveillance due to its angle and height being captured. Likewise, in a work of aerial action recognition [13], a dataset was developed, and multiple deep learning algorithms have been tried.

Nonetheless, the performance of algorithms applied on this dataset is deficient. The problem with such a dataset is that the features are not explored, and hence the deep learning algorithm could not learn properly for the classification of action. Recently, a few datasets have been published in aerial and drone surveillance and can be used for action recognition, such as in [36,37]. In the past few years, the use of a drone has been increasing, and it covers almost every field of surveillance, such as road safety and traffic analysis [38], crop monitoring [39], and border area surveillance [40]. Recent work has been published in this field, which uses drone surveillance using a deep learning action recognition model for search and rescue [41]. In the field of aerial action recognition, [42] proposed a disjoint multitasking approach for action detection in drone videos. However, the training sample is deficient and could not be a generalized framework for all.

These drone applications need models trained on an aerial dataset for analysis and surveillance to improve performance. Pre-trained object detection (OD) and action detection models are not useful in drones or aerial surveillance. The performances with diverse ground-level object detection tasks are popular for models like Mask R-CNN, RCNN [43], SPP-Net [44], Fast R-CNN [45], and Faster R-CNN [46]. These models were used to identify actions; however, their inference time is the main drawback for such two-stage OD algorithms. In the OD family, SSD [47] has performance comparable to other listed algorithms and lowers inference time to one of the best inference time algorithms such as YOLO (You Look Only Once). However, as shown in work [13], these algorithms are not sufficiently good for long-term and aerial videos. In some works, a separate module for detecting the human body and its extension with other models for detecting action is used [48]. Due to its feature and angle, it may be helpful to use this separate module for both tasks for aerial monitoring.

Table 1. Key research and dataset published in the field of drone surveillance-based human detection and action recognition task.

S.No	Title	Result	Comments
1.	Okutama-Action: An Aerial View Video dataset for Concurrent Human Action Detection [13]	0.18 mAP@0.50IOU	1. SSD-based object detection model is used for object detection and action recognition. 2. Dataset is proposed for drone-based action recognition, and this is one of the very first public datasets for drone-based action recognition to the best of my knowledge.
2.	Convolution Neural Networks for Aerial Multi-Label Pedestrian Detection [14]	0.28 mAP@0.50IOU	1. Okutama Dataset is used for multi-label pedestrian detection 2. Slight improvement in performance, however, it is far from real-time use.
3.	UAV-Gesture: A dataset for control and gesture recognition [37]	85% validation accuracy	1. In this article, a new dataset was proposed for drone-based gesture recognition. 2. However, as a single human was present in all frames of this dataset, it is not useful for multi-action gesture recognition.
4.	Drone Surveillance for Search and Rescue [36]	87% accuracy	1. This paper proposes a novel dataset for multi-label human detection and action recognition in drone surveillance. 2. However, this dataset is captured inside the campus with a similar background, and applying action recognition directly to this dataset can lead to inefficient real-time action recognition in disasters.
5.	Ucf roof-top-dataset [49]	76% accuracy	1. Dataset is published with a still camera fitted at the top. 2. Dataset is captured in similar background.
6.	Campus [50]		1. Another dataset proposed for drone-based surveillance. 2. Number of classes is one in this dataset, so comparatively easier for models to be trained. 3. Number of frames and variety in the dataset is not satisfactory.

3. Dataset

The dataset used for training and testing is described in this section. It also provides details of the dataset developed for the recognition of aerial action. The dataset is also compared with aerial or ground video datasets, which have been recently published.

3.1. Human Detection Dataset

Human detection is an essential sub-module of the proposed architecture. The proposed architecture of drone surveillance for action recognition requires humans to be detected precisely and requires an accurate human detection dataset. For our sub-module human detection, this work utilizes the dataset proposed in [38], in which the images are captured from the required height and angle. As a result, the dataset has more than 5000 humans annotated from different angles and heights. In addition, the dataset has a good variety in terms of actors, weather conditions, and the daytime.

3.2. Proposed Action Recognition Dataset

The second important module in the proposed architecture is action recognition. For this, an accurate dataset is required. Unfortunately, a recently published dataset for aerial action recognition has mostly been captured for multi-view action performed by more than one human, where recognizing a single action is inaccurate. For this, this paper proposes a

dataset required for training with five different classes. The process of the development of the dataset is as follows.

3.2.1. Dataset Collection

The dataset is collected in the form of video clips and has more than 30 actors performing different actions for our 480 videos. The dataset is collected at different times of day on varying days of the month. Actors wearing varying clothes on a different day of capture give a quality variation for deep learning models. The captured video has actors of different ages and gender. The videos are captured in various parts of India and mostly in a rural background, making them useful for search and rescue tasks.

3.2.2. Dataset Preprocessing

All video captured is then resized into the single-size video format without disturbing the aspect ratio of frames. Each video is then labeled into five different classes as required, especially for search and rescue. Data are also converted in the image form to test 2D ConvNet for action recognition. The frame was selected in the ratio of 1:5 in the image dataset for action recognition. However, the image dataset has only spatial features, which are sometimes not useful for classifying almost overlapping action such as human walking and human standing, which cannot be classified with only spatial features.

3.2.3. Summary of Proposed Dataset

The dataset contains a total of 480 clips having 14,400 frames. The dataset is labeled in five different classes manually, and the number of videos is equally divided between all five classes. The important detail of the proposed dataset is summarized in Table 2. One can see in this table that the UCF rooftop dataset has almost 50 video clips in each class. However, it was released with more than five classes. The proposed five actions for search and rescue in a natural disaster are sufficient for identifying the emergency. It has sufficient variation in the phase, orientation, camera movement, viewpoint, number of humans in the action, human sizes, and the backgrounds.

Table 2. Features of action classification dataset.

Class	Video Type	Background	No. of Videos in the Proposed Dataset	No. of Videos in UCF Rooftop Dataset
Fight	.MP4	Rural area, College campus	85	48
Lying	.MP4	Rural area, College campus	85	48
Shaking Hand	.MP4	Rural area, College campus	85	48
Waving Hand	.MP4	Rural area, College campus	85	48
Waving Both Hand	.MP4	Rural area, College campus	85	0

4. Methodology

The proposed approach for search and rescue uses drones as a tool for monitoring disaster-affected areas. This paper develops an end-to-end trainable system for faster motion feature modeling and action recognition from crowd surveillance videos. The architecture of the proposed methodology is explained in Figure 1. This approach combines two different modules. Hence, the architecture and its work are presented with the help of each module's mathematical background. Some necessary notations used in this paper are as follows: For the video clips, each video clip Xi = [c ∗ l∗ h ∗ w]. Where I = f (x, y) represents the frames of the video clip, c stands for the number of channels, l is for the number of the frame in each video clip, h and w represents the height and width of each frame.

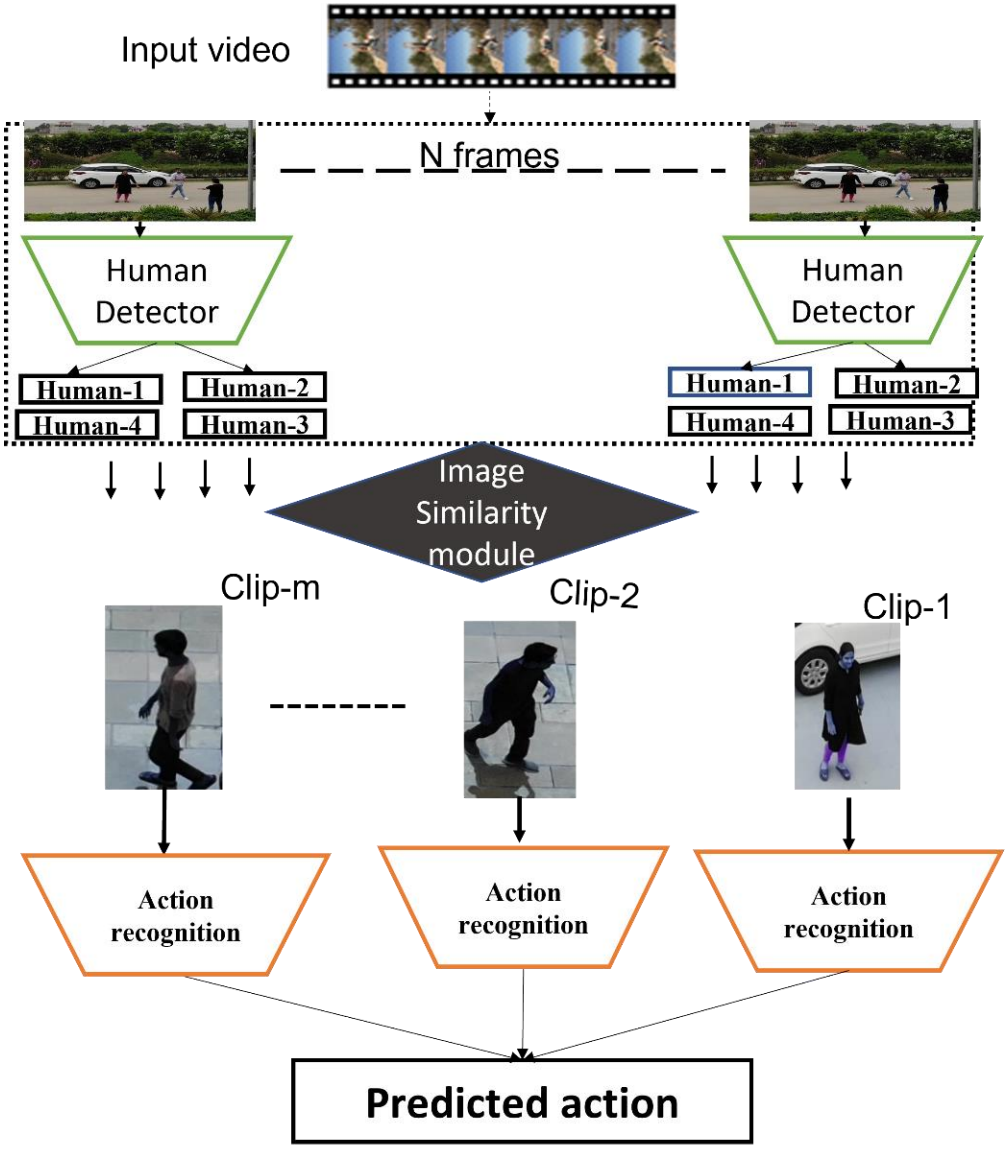

Figure 1. Flow diagram for proposed background-invariant search and rescue using a combination of action recognition and human detection model.

4.1. Faster Motion Feature Modeling (FMFM)

Multiple modules are interconnected and trained together for the accurate and faster motion modeling of temporal features. This combines deep learning model object detection and traditional computer vision algorithms to extract detected bounding boxes. First, an individual human's temporal feature model is extracted based on the image similarity from the original crowd monitoring video. For this, the initial step is detecting the human precisely. After the humans are detected accurately, the output image is passed to the second part of FMFM, which extracts the detected human, calculates image similarity, and patches most similar extracted images together in the form of video-clip. Thus, the whole process of FMFM is divided into two critical steps.

- Accurately detecting humans in the surveillance video is the primary step. For this, object detection techniques could help. The benchmark results of object detection techniques motivate us to apply them for aerial human detection. Figure 2 shows some sample of output images after applying the object detection models on drone images and it shows the models have accurately detected humans after training the models on the aerial surveillance image dataset using transfer learning.

- The second and most important step in FMFM is extracting the detected humans from the inference image of the object detection algorithm and combining this with the other extracted humans from a different frame with the highest similarity.

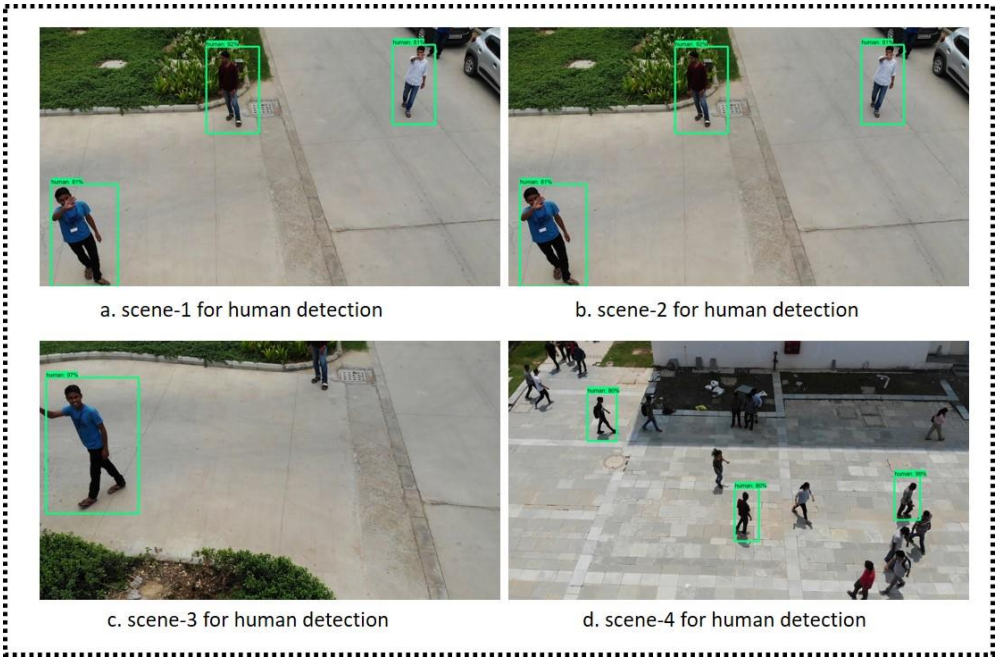

Figure 2. Sample images for human detection by human detection module, sub-image (**a**–**d**) represent the detected humans from different angles, backgrounds, and heights of the drone.

In this way, the crowd monitoring video dataset is converted into individual human action videos. Thus, each video has a sufficient human feature, which can be utilized later for accurate action recognition. Another advantage of FMFM is that it removes the background's effect, affecting the performance of the action recognition algorithm. Mostly, models are trained in a typical environment, and it fails when applied in a real-time environment or when the background changes.

As discussed above, the process of FMFM starts with detecting humans precisely, so here are some state-of-the-art object detection techniques that can be utilized for detecting humans in aerial surveillance images and videos.

- Faster R-CNN: The original paper of Faster R-CNN [46], the model with less inference time for object detection, was proposed based on a region-based approach. In this, fewer region proposals are generated than other region-based object detection category algorithms. Initially, the VGG-16 network was used as a feature extractor for region proposals in the Faster RCNN network. However, over time, some advanced models, like inception and MobileNet, were used to implement Faster R-CNN. Here, in our experiment, This module uses Faster RCNN with an inception network as a feature extractor.
- SSD (Single-shot detection): SSD was originally proposed in [47], and it uses multiple-layer features of feature extractor network VGG16 for detecting the objects. In this approach, classification and localization are both done with a single network. In our proposed model for object detection, inception and MobileNet's base architecture are used as the backbone.

Basically, object detection is a combination of object localization and object classification in an image I = f(h, w). Applying object detection on a video stream or video-clip $Xi = I^{c*l*h*w}$ requires applying the same process on each frame of the video clip.

$$Yi = O(I) \quad (1)$$

where Yi is the output label, i.e., human, car, truck, bus, for a good quality object detection model, two different types of loss, localization loss and classification loss, need to be optimized. Therefore, the localization loss is defined as Lloc (y, y′) and classification loss as Lclass (y, y′), in which y is the ground truth object class, and y′ denotes the predicted object class.

$$\mathrm{L}(\mathrm{y},\ \mathrm{y}') = \mathrm{Lloc}(\mathrm{y},\ \mathrm{y}') + \mathrm{Lclass}(\mathrm{y},\ \mathrm{y}') \tag{2}$$

4.2. Accurate Action Classification (AAR)

This paper approaches the search and rescue task for humans as a multi-class activity recognition problem in drone surveillance. Usually, in drone surveillance, the height and angle of surveillance change with time, making it more difficult. Identifying a specific action from the image captured where multiple humans are present and possibly all doing different activities is quite challenging and requires detecting each action individually. Therefore, the proposed architecture uses two different modules, FMFM and AAR, which accurately detect each action. However, in drone video surveillance, action recognition could be performed with the help of spatial features. The spatial features of humans could be combined with motion features to classify the actions accurately. In both categories, various experiments have been performed in this study with the existing models and the combination of different architecture such as 2D ConvNet with RNN and 2D ConvNet with Time-Distributed RNN and with the proposed architecture using 3D ConvNet. However, for recognizing few actions required for emergency identification in disaster, motion features are required, and hence this paper finds models that use spatial and motion features together are more suitable for such tasks. For example, in the case of finding humans through their actions, waving hands is a primary action, requiring spatial and temporal features. The mathematical modeling of action recognition is as follows:

$$\mathrm{Zi} = \mathrm{A}\ (\mathrm{I}) \tag{3}$$

$$\mathrm{Zi} = \mathrm{A}\left(\mathrm{I}^{\mathrm{C}*\mathrm{l}*\mathrm{h}*\mathrm{w}}\right) \tag{4}$$

where Zi represents the output class of the action recognition model, I represent the image, and $\mathrm{I}^{\mathrm{c}*\mathrm{l}*\mathrm{h}*\mathrm{w}}$ represents the input video for action recognition module.

$$\mathrm{Zi} = \sum_{\mathrm{r}\ =2}^{\mathrm{w}} \mathrm{A}(\mathrm{Ir})) + \mathrm{T}\left(\mathrm{A}\left(\mathrm{I}^{\mathrm{c}*\mathrm{l}*\mathrm{h}*\mathrm{w}}\right)\right) \tag{5}$$

where function A represents the action recognition based on 2D spatial features, and T function represents the action classification feature extraction based on motion features. Equation (5) represents the working of CNN + RNN, and CNN + Time distributed RNN models used in our experimentation, where two different networks are used to extract features parallel in the end, both the features are combined for action classification.

4.2.1. Action Classification with Spatial Features

Human actions such as lying, sitting, fighting, shaking hands could be easily recognized by the image's spatial features. Image classification models can recognize such features. Some important 2D ConvNet used in this paper is as follows:

- Resnet50: This is a 50-layer deep convolution neural network that uses the residual function. It starts from 7×7 convolution with stride two and max pool kernel of size 3×3 with stride 2. After that, at four different stages, it uses the residual function of sizes 3, 4, 6, and 3. Thus, each residual function contains three convolution networks of sizes 1×1, 3×3, and 1×1 [51].
- Resnet152: In this, the network depth was 152 layers with the residual function. It also starts with a convolutional layer of 7×7, and all other convolution layers are patched in the form of residual function as in Resnet50 [51].

- VGG16: This is a type of convolution layer with 16 layers. In this network, 13 convolution layers are present with three fully connected layers [52].

4.2.2. Action Classification with Temporal Features

Multiple architectures were designed and trained for action recognition with spatio-temporal features. Spatio-temporal features are captured through different networks such as 2D ConvNet, RNN, and time distributed RNN. The architecture of the models was as follows:

- 2D ConvNet with RNN: In this research, the experiments were performed with a five-layer 2D ConvNet working in parallel with three layers RNN combined with three dense layers in the end for feature classification. Two-dimensional ConvNet extracts spatial features, and RNN networks extract the motion features from the sequence of frames. Extracted features are then passed to the network's dense layer after flattening it into a single dimension vector. Both networks working in parallel are merged into a single network with the concatenate feature available in Keras.289
- 2D ConvNet with time distributed RNN: this network consists of 5-layer 2D ConvNet with three layers of time distributed RNN layer and one RNN layer working in parallel for spatial and motion feature extraction. Both the networks are then merged for overall feature classification with three dense layers finally.
- Proposed architecture of 3D ConvNet: our proposed network for action recognition uses the modified architecture of VGG16 with 12 3D ConvNet and two dense layers in the end. Convolution layers are stacked in the fashion as it appears in VGG16 with $3 \times 3 \times 3$ convolution filters in the first layer having 64 kernels. The detailed structure of the proposed network is represented in Table 3. The relu activation function is used at each layer except the last dense layer, where the softmax activation function is applied.

Table 3. Detail of proposed convolution neural network for action recognition.

Layer	Kernel Size	Output Shape
3D Conv	$3 \times 3 \times 3$	(None, 32, 32, 12, 64)
3D Conv	$3 \times 3 \times 3$	(None, 32, 32, 12, 128)
3D Conv	$3 \times 3 \times 3$	(None, 32, 32, 12, 128)
Maxpool	$2 \times 2 \times 2$	(None, 16, 16, 6, 128)
3D Conv	$3 \times 3 \times 3$	(None, 16, 16, 6, 256)
3D Conv	$3 \times 3 \times 3$	(None, 16, 16, 6, 256)
3D Conv	$3 \times 3 \times 3$	(None, 16, 16, 6, 256)
Maxpool	$2 \times 2 \times 2$	(None, 8, 8, 3, 256)
3D Conv	$3 \times 3 \times 3$	(None, 8, 8, 3, 512)
3D Conv	$3 \times 3 \times 3$	(None, 8, 8, 3, 512)
3D Conv	$3 \times 3 \times 3$	(None, 8, 8, 3, 512)
Maxpool	$2 \times 2 \times 2$	(None, 4, 4, 2, 512)
3D Conv	$3 \times 3 \times 3$	(None, 4, 4, 2, 512)
3D Conv	$3 \times 3 \times 3$	(None, 4, 4, 2, 512)
3D Conv	$3 \times 3 \times 3$	(None, 4, 4, 2, 512)
Maxpool	$2 \times 2 \times 2$	(None,2, 2, 1, 512)
Flatten	None	(Non, 2048)
Fully connected	None	(None, 512)
Fully connected	None	(None, 5)

4.3. Proposed Architecture and Its Operation

The proposed operational architecture contains two different stages, i.e., FMFM and AAR as explained above. In this, in the beginning, humans are detected in surveillance videos, and the output of the human detection module is then passed to the function written for cropping the bounding boxes predicted. It uses human coordinates detected by a human detection module for a single or multiple humans in a surveillance image.

Furthermore, the cropped image function's output is passed to another module, which combines the cropped images based on similarity. For the similarity of images, multiple techniques were tested and based on the best outcome with the mean square error (mse) result, and the proposed module uses mse for stacking.

$$X_i = \text{mse}(i_1,\ i_2,\ i_3,\ \ldots\ldots i_n) \tag{6}$$

where V_i is the video clip formed with mse (mean square error) parameter. mse is used here to group mot similar images together as per the pixel values of images:

$$(i_1,\ i_2,\ i_3,\ \ldots\ldots i_{60}) = C_{crop}(I_1,\ I_2,\ I_3,\ \ldots\ldots I_{60}) \tag{7}$$

where $I_1,\ I_2, I_3 \ldots\ldots I_{60}$ are the individual output image from the object detection module. Each detected human is cropped by the function, C_{crop} and $i_1, i_2, i_3,\ \ldots\ldots i_n$ are the different instances of humans cropped.

In this, the proposed architecture uses a patch human detected output of 60 consecutive frames together for stacking it into individual action recognition video clips. Finally, the output individual action video clips are passed through the proposed action recognition sub module for final action recognition.

The overall objective of the proposed approach is to minimize the unified loss represented by L_{BackI}, which is a combination of object detection loss $L(y,\ y')$, and mean square error (mse) for patching action classification loss L_{Aclass}. The detailed equation of loss is explained through Equation 8. In this, y and y' represent the output class for human detection. Human and non-human are the class labels for this. Overall, our final output is based on the 5-action class, where z represents the ground truth labels and z' is for the predicted class.

$$L_{BackI}(z,\ z') = L(y,\ y') + \text{mse}(f1(h,w), f2(h,w)) + L_{Aclass} \tag{8}$$

5. Experiments and Results

The result of the experiments was evaluated with the proposed dataset using the metric of action classification accuracy. Furthermore, extensive experiments were performed with various 2-D and 3-D models, which deal with the spatial and temporal features of input videos in the dataset. This section covers the details of all the experiments and reports the result in the most suitable format.

5.1. Experimental Setup

The experiment is set up on the 6 GB RAM, 4 GB NVIDIA-GPU workstations. Python 3 which is an opensource software is used with its OpenCV package for preprocessing and post-processing the dataset. The tensorFlow environment is used on top of Python3 as the base for all the experiments performed in this paper. In addition, various other Python packages, such as Keras, tensorboard, etc., were used throughout the experiments.

5.2. Experiments

The experiments were performed on the dataset described in Section 3, which consists of two different datasets, one for human detection and the other for action recognition proposed in this paper. At the primary level, experiments were performed with object detection models for recognizing the action in drone videos. After some preprocessing, it is converted into five-class action recognition with the dataset for human detection, and three different object detection models were trained using transfer learning. Experimental results with object detection model, Faster RCNN, SSD Inception, and SSD MobileNet for action recognition motivate us to develop an architecture where the features are exposed for classification. Therefore, this paper proposed a novel architecture for this. The proposed architecture was tested with the help of two different trainable networks, FMFM and AAR. For the first module, three different human detection modules were trained with more

than 20,000 steps and their best-tuned hyperparameters. In this phase, cropping Haman and stacking the maximum matched extracted human, which is also be termed as motion feature modeling of the individual human, was performed to optimize different image matching algorithms.

The second stage of our experimentation was designing and testing a novel action recognition model that could recognize drone surveillance videos' human action. For this, three different models, with the help of transfer learning, were trained on our dataset. Moreover, as the learning of 2D models is not sufficient for accurately classifying the action, two different state-of-the-art models were also trained on our dataset (CNN + RNN, CNN + TimeDLSTM). These experiments motivated us to design a slightly complex but accurate model for classifying the models based on the video's spatio-temporal features. All these models were trained and were tuned with their best hyperparameter values.

5.2.1. Training and Testing

In this paper, the result was obtained and reported by training three models for object detection for action recognition and the same three models for human detection for the proposed architecture. For the second module, AAR, three 2D ConvNets were trained on our dataset, also, including proposed network state-of-the-art models that were trained and optimized for the best hyperparameter values. Twelve different networks were trained and compared based on the evaluation parameters such as mAP (mean average precision) for object detection models and validation accuracy for action classification models. For the training and testing of models, data were divided into a 7:3 ratio. Each model was trained up to the saturation point.

5.2.2. Results

The proposed action recognition module uses a convolution neural network feature for feature extraction in a video frame. Starting from the object detection models applied, Table 4 represents the result, and it shows that the performance of SSD model is comparable to Faster RCNN with lesser inference time. Here, Faster RCNN stands for Faster Region Based Convolution Neural Network model in the family of object detection. While, two variant of SSD (Single shot detection) are used, ie. With Inception as backbone, and MobileNet variant. Performance is compared based on the primary performance measure metric mAP stands for mean average precision, and then inference time with unit milli seconds is used. In contrast, the recurrent neural network works with multiple video frames to extract and utilize the temporal feature to classify video into a class. For action recognition models trained on our dataset, the result of each model is reported in Tables 5 and 6. Table 5 summarizes the experimentation result with the 2D ConvNet trained on our dataset using transfer learning with three different models, ie. Resnet 50, resnet 152, and VGG16. The result obtained in this paper is compared with the state-of-the-result in Table 7. Moreover, Table 7 reports the result of all the models applied to our dataset based on parameters training accuracy, training loss, validation accuracy, and validation loss. Okutama is a standard dataset published for aerial human detection and action recognition with the benchmark result 0.18 mAP as shown in Table 7. While, in some other work listed in Table 7, multilevel pedestrian detection (PD) has been attempted with 0.28 mAP. Convolution neural network (CNN) along with recurrent neural network (RNN) and time distributed long short term memory unit (LSTM) has also been used and we have compared their result in Table 7. Figures 3 and 4 represent the loss and accuracy of the proposed action recognition model trained on our developed dataset. The proposed architecture's incremental testing was performed in another level of testing, and each module was tested for the outcome as action recognition. The result of incremental module testing is reported in Table 8.

Table 4. Accuracy and inference time comparison of object detection models for human detection trained on our human detection dataset.

Object Detection Model	mAP	Speed (milliseconds)
Faster RCNN Inception	0.833	52
SSD Inception	0.848	42
SSD Mobilenet	0.796	30

Table 5. Performance of 2D convolution model applied on our dataset for action recognition.

Network	Validation Accuracy
Resnet 50	0.65
Resnet 152	0.61
VGG 16	0.91

Table 6. Detailed result comparison of models applied on the proposed dataset using transfer learning with proposed model performance.

Network	Accuracy	Loss	Validation Accuracy	Validation Loss
3D Resnet [10]	0.39 mAP	0.33	0.65	1.45
CNN + RNN	0.29 mAP	0.33	0.45	1.75
CNN + TimeDLSTM	0.85	0.33	0.53	1.33
Proposed model	0.90	0.25	0.85	0.75

Table 7. Comparison of aerial action recognition models.

Network	Validation Accuracy
Okutama SSD [13]	0.18 mAP
CNN for multilevel PD [53]	0.28 mAP
3D Resnet [10]	0.65
MOD-20 [54]	0.74
CNN + RNN (on our dataset)	0.45
CNN + TimeDLSTM (on our dataset)	0.53
The proposed model (on our dataset)	0.85

Table 8. Incremental analysis of modules directly applied for action recognition in the proposed action dataset.

Network	Validation Accuracy
Faster RCNN Inception	0.39 mAP
SSD Inception	0.29 mAP
Proposed model for action recognition	0.85
Combined architecture proposed for action recognition	0.90

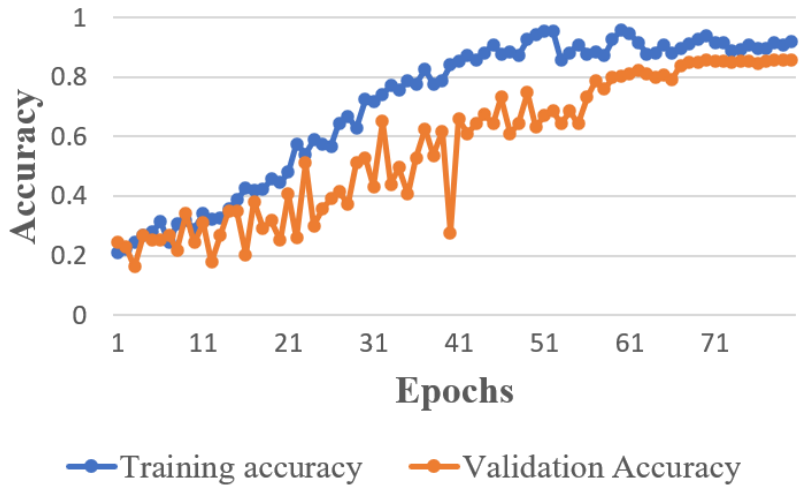

Figure 3. Result of the proposed model for training and validation accuracy with the proposed dataset.

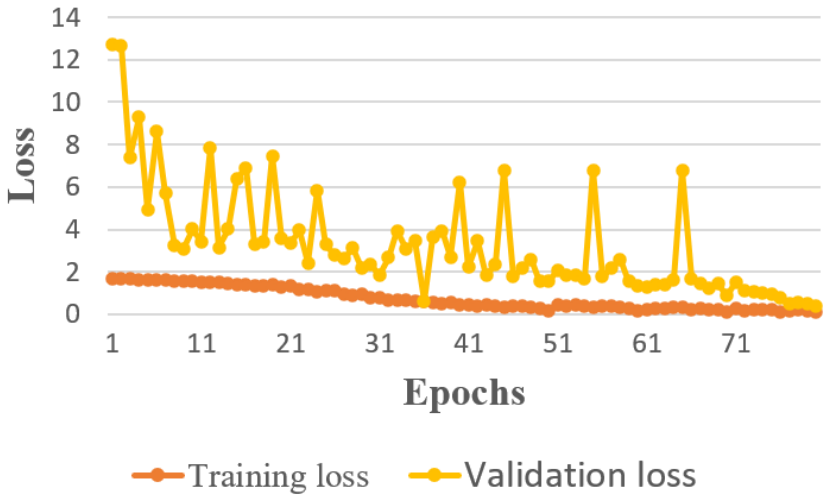

Figure 4. Result of the proposed model for training and validation loss with the proposed dataset.

6. Discussion

The proposed architecture was tested at various levels, and one of the most important analyses we performed was through incremental module testing. Although in some pre-various work, object detection models have been directly used to recognize action, we also applied it directly to recognize the action. Table 8 shows the performance of object detection models applied for action recognition. However, it achieves only 0.39 mAP with the state-of-the-art model faster RCNN. However, when it is applied for human detection, it can detect the human with a 0.833 mAP value; the comparison of various human detection models trained and tested on the proposed dataset is represented in Table 4. The second module proposed an action recognition model and was also tested independently on the proposed action recognition video dataset. The maximum accuracy we achieved with the proposed model was 0.85, which was validation accuracy with a training accuracy of 0.89. The detailed comparative result is represented in Tables 7 and 8. Finally, this paper proposes an accurate and faster 3D temporal motion modeling architecture, and this was applied to the input video. While testing the architecture with the validation video, the architecture achieved maximum accuracy with 0.90 validation accuracy and is represented in Table 8. However, these performances have certain limitations: as the height of the drone will increase, the human feature changes, especially in drone footage captured from above. These experiments were performed on the drone footage captured from our DJI Mavic Pro drone, with the camera having a 60 frame per second rate. Images and videos were

captured from 10 to 50 m height. Some important failure cases arose when we tested with the dataset published in the aerial image of the Okutama dataset since the height from where those images were captured was more than 60 m. Our experimental result, and the incremental analysis of each module, observed that using both proposed modules (FMFM and AAR) together could achieve higher accuracy than any other combination.

7. Conclusions

In this paper, a novel architecture for motion feature modeling is proposed and invariant to background effects that are useful for disaster and search and rescue applications. The proposed architecture uses two modules: FMFM, which models the temporal feature and creates background invariant clips which work as input for the second module, AAR. Also, a specific dataset is proposed and is validated with various state-of-the-art models. Moreover, an accurate model for temporal action recognition is suggested in this paper, which outperforms all the action recognition models applied to aerial action recognition. The performance of the overall architecture is increased compared to the performance of the proposed action recognition algorithm. This happens because when both modules work together, the AAR module will obtain the video clip with more specific human action features, and hence the performance will increase. Overall performance of the proposed architecture is more than 90%, which sets a new benchmark in aerial action recognition.

Author Contributions: Conceptualization: B.M. and D.G. Methodology: B.M., K.K. and P.N. Software: B.M. and V.K.M. Validation: B.M., K.K., D.G. and V.K.M. Formal analysis: P.N. Investigation: K.K. Data curation: B.M. Writing—original draft preparation: B.M. Writing—review and editing: P.N. and D.G. Visualization, B.M. and V.K.M. Supervision, D.G. and P.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Feichtenhofer, C.; Pinz, A.; Wildes, R.P. Spatiotemporal multiplier networks for video action438recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4768–4777.
2. Sun, S.; Kuang, Z.; Sheng, L.; Ouyang, W.; Zhang, W. Optical flow guided feature: A fast and robust motion representation for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1390–1399.
3. Wang, Y.; Long, M.; Wang, J.; Yu, P.S. Spatiotemporal pyramid network for video action recognition. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1529–1538.
4. Li, J.; Liu, X.; Zhang, M.; Wang, D. Spatio-temporal deformable 3d convnets with attention for action recognition. *Pattern Recognit.* **2020**, *98*, 107037. [CrossRef]
5. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, P.; et al. The kinetics human action video dataset. *arXiv* **2017**, arXiv:1705.069502017.
6. Li, A.; Thotakuri, M.; Ross, D.A.; Carreira, J.; Vostrikov, A.; Zisserman, A. The ava-kinetics localized human actions video dataset. *arXiv* **2020**, arXiv:2005.00214.
7. Materzynska, J.; Berger, G.; Bax, I.; Memisevic, R. The jester dataset: A large-scale video454dataset of human gestures. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, Seoul, Korea, 27–28 October 2019.
8. Wang, L.; Xiong, Y.; Wang, Z.; Qiao, Y.; Lin, D.; Tang, X.; Van Gool, L. Temporal segment networks: Towards good practices for deep action recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 20–36.
9. Ji, S.; Xu, W.; Yang, M.; Yu, K. 3D convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 221–231. [CrossRef] [PubMed]
10. Yang, C.; Xu, Y.; Shi, J.; Dai, B.; Zhou, B. Temporal pyramid network for action recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 591–600.

11. Shi, L.; Zhang, Y.; Cheng, J.; Lu, H. Skeleton-based action recognition with multi-stream465adaptive graph convolutional networks. *IEEE Trans. Image Process.* **2020**, *29*, 9532–9545. [CrossRef]
12. Perera, A.G.; Law, Y.W.; Chahl, J. Drone-Action: An Outdoor Recorded Drone Video Dataset for Action Recognition. *Drones* **2019**, *3*, 82. [CrossRef]
13. Barekatain, M.; Martí, M.; Shih, H.F.; Murray, S.; Nakayama, K.; Matsuo, Y.; Prendinger, H. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 28–35.
14. Kim, S.J.; Lim, G.J. Drone-aided border surveillance with an electrification line battery charging system. *J. Intell. Robot. Syst.* **2018**, *92*, 657–670. [CrossRef]
15. Li, M.; O'Grady, M.; Gu, X.; Alawlaqi, M.A.; O'Hare, G.; Wan, J. Time-bounded activity recognition for ambient assisted living. IEEE Trans. *Emerg. Top. Comput.* **2018**, *9*, 471–483.
16. Robicquet, A.; Sadeghian, A.; Alahi, A.; Savarese, S. Learning social etiquette: Humantrajectory understanding in crowded scenes. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 549–565.
17. Sevilla-Lara, L.; Liao, Y.; Güney, F.; Jampani, V.; Geiger, A.; Black, M.J. On the integration of optical flow and action recognition. In *Proceedings of the German Conference on Pattern Recognition*; Springer: Berlin, Germany, 2018; pp. 281–297.
18. Zhu, Y.; Lan, Z.; Newsam, S.; Hauptmann, A. Hidden two-stream convolutional networks for action recognition. In Proceedings of the Asian Conference on Computer Vision, Perth, Australia, 2–6 December 2018; Springer: Berlin/Heidelberg, Germany, 2018; pp. 363–378.
19. Baccouche, M.; Mamalet, F.; Wolf, C.; Garcia, C.; Baskurt, A. Sequential deep learning for human action recognition. In Proceedings of the International Workshop on Human Behavior Understanding, Amsterdam, The Netherlands, 16 November 2011; Springer: Berlin/Heidelberg, Germany, 2011; pp. 29–39.
20. Xia, L.; Chen, C.C.; Aggarwal, J.K. View invariant human action recognition using histograms of 3d joints. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 20–27.
21. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Spatio-temporal lstm with trust gates for 3d human action recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 816–833.
22. Danafar, S.; Gheissari, N. Action recognition for surveillance applications using optic flow and SVM. In Proceedings of the Asian Conference on Computer Vision, Tokyo, Japan, 18–22 November 2007; Springer: Berlin/Heidelberg, Germany, 2007; pp. 457–466.
23. Ohn-Bar, E.; Trivedi, M. Joint angles similarities and HOG2 for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, USA, 23–28 June 2013; pp. 465–470.
24. Bloom, V.; Makris, D.; Argyriou, V. G3D: A gaming action dataset and real time action recognition evaluation framework. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 7–12.
25. Gall, J.; Yao, A.; Razavi, N.; Van Gool, L.; Lempitsky, V. Hough forests for object detection, tracking, and action recognition. IEEE Trans. *Pattern Anal. Mach. Intell.* **2011**, *33*, 2188–2202. [CrossRef]
26. Soomro, K.; Zamir, A.R.; Shah, M. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv* **2012**, arXiv:1212.04022012.
27. Hou, R.; Chen, C.; Shah, M. Tube convolutional neural network (T-CNN) for action detection in videos. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5822–5831.
28. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 568–576.
29. Materzynska, J.; Xiao, T.; Herzig, R.; Xu, H.; Wang, X.; Darrell, T. Something-else: Compositional action recognition with spatial-temporal interaction networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1049–1059.
30. Dai, C.; Liu, X.; Lai, J. Human action recognition using two-stream attention based LSTM networks. *Appl. Soft Comput.* **2020**, *86*, 105820. [CrossRef]
31. Li, C.; Li, S.; Gao, Y.; Zhang, X.; Li, W. A Two-stream Neural Network for Pose-based Hand Gesture Recognition. *arXiv* **2021**, arXiv:2101.089262021.
32. Luvizon, D.C.; Picard, D.; Tabia, H. 2d/3d pose estimation and action recognition using multitask deep learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5137–5146.
33. Liu, M.; Yuan, J. Recognizing human actions as the evolution of pose estimation maps. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1159–1168.
34. Shahroudy, A.; Liu, J.; Ng, T.T.; Wang, G. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1010–1019.
35. Chen, C.; Jafari, R.; Kehtarnavaz, N. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP), Québec City, QC, Canada, 27–30 September 2015; pp. 168–172.

36. Mishra, B.; Garg, D.; Narang, P.; Mishra, V. Drone-surveillance for search and rescue in natural disaster. *Comput. Commun.* **2020**, *156*, 1–10. [CrossRef]
37. Perera, A.G.; Wei Law, Y.; Chahl, J. UAV-GESTURE: A dataset for UAV control and gesture recognition. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
38. Salvo, G.; Caruso, L.; Scordo, A. Urban traffic analysis through an UAV. *Procedia-Soc. Behav. Sci.* **2014**, *111*, 1083–1091. [CrossRef]
39. Mogili, U.R.; Deepak, B. Review on application of drone systems in precision agriculture. *Procedia Comput. Sci.* **2018**, *133*, 502–509. [CrossRef]
40. Kim, S.J.; Lim, G.J. A hybrid battery charging approach for drone-aided border surveillance scheduling. *Drones* **2018**, *2*, 38. [CrossRef]
41. Mishra, B.; Garg, D.; Narang, P.; Mishra, V. A hybrid approach for search and rescue using 3DCNN and PSO. *Neural Comput. Appl.* **2021**, *33*, 10813–10827. [CrossRef]
42. Sultani, W.; Shah, M. Human Action Recognition in Drone Videos using a Few Aerial Training Examples. *arXiv* **2021**, arXiv:1910.100272019.
43. Uijlings, J.R.; Van De Sande, K.E.; Gevers, T.; Smeulders, A.W. Selective search for object recognition. *Int. J. Comput. Vis.* **2013**, *104*, 154–171. [CrossRef]
44. Purkait, P.; Zhao, C.; Zach, C. SPP-Net: Deep absolute pose regression with synthetic views. *arXiv* **2017**, arXiv:1712.034522017.
45. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 5–9 June 2015; pp. 1440–1448.
46. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.
47. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the European conference on computer vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
48. Chakraborty, B.; Rudovic, O.; Gonzalez, J. View-invariant human-body detection with extension to human action recognition using component-wise HMM of body parts. In Proceedings of the 2008 8th IEEE International Conference on Automatic Face & Gesture Recognition, Amsterdam, The Netherlands, 17–19 September 2008; pp. 1–6.
49. U. of Central Florida.Ucf-arg Dataset. 2020. Available online: https://www.crcv.ucf.edu/data/UCF-ARG.php (accessed on 3 March 2019).
50. Li, Q.; Gravina, R.; Li, Y.; Alsamhi, S.H.; Sun, F.; Fortino, G. Multi-user activity recognition: Challenges and opportunities. *Inf. Fusion* **2020**, *63*, 121–135. [CrossRef]
51. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
52. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.15562014.
53. Soleimani, A.; Nasrabadi, N.M. Convolutional neural networks for aerial multi-label pedestrian detection. In Proceedings of the 2018 21st International Conference on Information Fusion (FUSION), Cambridge, UK, 10–13 July 2018; pp. 1005–1010.
54. Perera, A.G.; Law, Y.W.; Ogunwa, T.T.; Chahl, J. A Multiviewpoint Outdoor Dataset for Human Action Recognition. *IEEE Trans. Hum. Mach. Syst.* **2020**, *50*, 405–413. [CrossRef]