drones

MDPI

Article

Drone-Action: An Outdoor Recorded Drone Video Dataset for Action Recognition

Asanka G. Perera [1,*], Yee Wei Law [1] and Javaan Chahl [1,2]

[1] School of Engineering, University of South Australia, Adelaide SA 5095, Australia; yeewei.law@unisa.edu.au (Y.W.L.); Javaan.Chahl@unisa.edu.au (J.C.)
[2] Joint and Operations Analysis Division, Defence Science and Technology Group, Melbourne, VIC 3207, Australia
* Correspondence: asanka.perera@mymail.unisa.edu.au

Received: 6 November 2019; Accepted: 24 November 2019; Published: 28 November 2019

Abstract: Aerial human action recognition is an emerging topic in drone applications. Commercial drone platforms capable of detecting basic human actions such as hand gestures have been developed. However, a limited number of aerial video datasets are available to support increased research into aerial human action analysis. Most of the datasets are confined to indoor scenes or object tracking and many outdoor datasets do not have sufficient human body details to apply state-of-the-art machine learning techniques. To fill this gap and enable research in wider application areas, we present an action recognition dataset recorded in an outdoor setting. A free flying drone was used to record 13 dynamic human actions. The dataset contains 240 high-definition video clips consisting of 66,919 frames. All of the videos were recorded from low-altitude and at low speed to capture the maximum human pose details with relatively high resolution. This dataset should be useful to many research areas, including action recognition, surveillance, situational awareness, and gait analysis. To test the dataset, we evaluated the dataset with a pose-based convolutional neural network (P-CNN) and high-level pose feature (HLPF) descriptors. The overall baseline action recognition accuracy calculated using P-CNN was 75.92%.

Keywords: drone; dataset; human action recognition; aerial video analysis; P-CNN

1. Introduction

Drones or Unmanned aerial vehicles (UAVs) are increasingly popular due to their affordability and applicability in numerous commercial applications. Some popular application areas of drones are photogrammetry [1], agriculture [2], crowd monitoring [3], sports activity recording [3], parcel delivery [3], and search and rescue [4,5]. Recently, these areas have experienced rapid advances through convergence of technologies [6].

For research in these areas, a series of datasets have been released in the past few years. These datasets cover multiple research disciplines but mainly in the security, industrial, and agricultural sectors. Examples of such application-specific drone datasets include datasets for object detection [7,8], datasets for vehicle trajectory estimation [9,10], datasets for object tracking [11,12], datasets for human action recognition [13–16], datasets for gesture recognition [17–19], datasets for face recognition [20,21], a dataset for fault detection in photovoltaic plants [22], datasets for geographic information system [23,24], and datasets for agriculture [25,26].

However, limited research has been done on human–drone interactions and associated applications. Although some commercial drones can recognize gestures, the capabilities rely on high-quality videos captured at low altitudes and are limited to a small number of simple gestures. Going beyond gestures, enabling a drone to recognize general human actions is far more challenging.

A limited number of studies have been conducted to understand the sophisticated human body movements recorded from a drone [16]. Reasons for the slow progress in human action recognition in aerial videos include:

- Human action recognition is inherently a challenging problem. Most of the action recognition studies are focused on popular video datasets that consist mostly of ground-level videos [27]. Even when using high quality videos, the state of the art in action recognition is still fallible [28].
- The quality of aerial videos is often marred by a lack of image details, perspective distortion, occlusions, and camera movements.
- Many drone applications require online rather than offline processing, but the resource constraints of embedded hardware platforms limit the choice and complexity of action recognition algorithms.
- There are not enough relevant video datasets to help train algorithms for action recognition in the aerial domain. Currently available aerial video datasets are mostly limited to object tracking. While there are some datasets supporting research in aerial action recognition [16], they are limited.

Based on the above, we reason that action recognition using aerial videos is less studied than general human action recognition. Recognizing the importance of datasets for action recognition research in the aerial domain, we present in this study, a new full high-definition (FHD) video dataset with rich human details, that was recorded on a drone in a controlled manner.

The work presented here is specifically focused on providing accurate and rich human body details in aerial videos recorded from a moving platform. State-of-the-art action recognition techniques mostly rely on high-resolution human videos to achieve high accuracy, typically by leveraging by computationally intensive techniques. However, many real-life aerial videos suffer from low resolution and undesired camera movement resulting from the apparatus. We believe that current limitations of aerial camera systems should not deprive action recognition research of high-quality and highly detailed aerial videos.

The dataset presented here consists of 13 action classes recorded at FHD (1920×1080 resolution) from a quadrotor drone flying slowly at low altitude. Some actions were recorded while the drone was *hovering* (e.g., kicking, stabbing, and punching), and others were recorded while the drone was *following* the subject (e.g., walking, jogging, and running). A hovering camera should not be confused with a stationary camera. A hovering camera can drift in any direction due to platform mobility and susceptibility to wind gusts. All of the videos were recorded in a way that preserved as much body surface area as possible. This dataset was designed to support research related to search and rescue, situational awareness, surveillance, pose estimation, and action recognition. We assume that in most practical missions, a drone operator or an autonomous drone follows these general rules:

- Avoid flying at overly low altitude, which is hazardous to humans and equipment;
- Avoid flying at overly high altitude, so as to maintain sufficient image resolution;
- Avoid high-speed flying, and therefore, motion blur;
- Hover to acquire more details of interesting scenes;
- Record human subjects from a viewpoint that gives minimum perspective distortion.

Our dataset was created assuming the guidelines above were followed.

For testing the dataset, a pose-based convolutional neural network (P-CNN) [29] was used as the baseline action recognition algorithm. P-CNN utilizes the CNN features of body parts extracted using the estimated pose. The dataset was also evaluated with high-level pose features (HLPF) [30] which recognize action classes based on the temporal relationships of body joints and their variations. The baseline accuracy and the experiment details have been compared with recently published human activity datasets.

The rest of this article is organized as follows. Section 2 discusses closely related work on popular action recognition datasets, aerial video datasets, and their limitations in terms of suitability for aerial human action recognition. Section 3 describes the steps involved in preparing the dataset,

and compares the dataset with other recently published video datasets. Section 4 reports experimental results. A discussion of issues and potential improvements is presented in Section 5. Section 6 concludes. Throughout this manuscript we will use the noun "drone" synonymously with the acronym UAV which describes an aircraft in terms of not being manned.

2. Related Work

A number of video datasets have been published for general human action recognition and aerial human action recognition. Here, we briefly discuss some studies related to our work.

Human action recognition datasets (mostly focused on ground videos): KTH [31] and Weizmann [32] were the most popular early action recognition datasets, and they were recorded in a controlled setting with a static background. These datasets helped to progress action recognition research. UCF101 [33] was one of the largest and most diverse datasets in this category. It was created with 13,320 YouTube videos belonging to 101 action classes. JHMDB [30] and Penn action [34] are another two notable action datasets created using YouTube videos. These datasets come with human pose and action annotations.

The recently made action recognition datasets are massive in terms of their numbers of videos and action classes. Sports-1M [35] and ActivityNet [36] were two datasets introduced under this category. These datasets do not provide spatial localization details of human subjects. For example, Sports-1M videos are annotated automatically using YouTube topics and are, therefore, weakly labeled [37]. The most recent additions to this category were YouTube-8M [38] Kinetics [39], and HACS [40]. The Kinetics dataset consisted of 400, 600, and 700 action classes in its successive releases. HACS dataset was created with 1.55 million annotated clips across 200 classes. The action classes of both datasets cover a diverse range of daily human actions collected from YouTube videos. Diversified large datasets are useful for pre-training action recognition networks on them and enhancing networks' performances on small datasets.

Aerial datasets: The available aerial video datasets can be broadly categorized as follows based on their intended applications.

- **Object detection and tracking datasets:** There has been a surge of interest in aerial object detection and tracking studies. They are mainly focused on vehicle and human detection and tracking. VisDrone [11] is the largest object detection and tracking dataset in this category. This dataset covers various urban and suburban areas with a diverse range of aerial objects (it includes 2.5 million annotated instances). The dataset has been arranged into four tracks for object detection in images/videos and single/multi object tracking. The similar but relatively smaller UAVDT dataset [41] contains 80 thousand annotated instances recorded from low altitude drones. UAV123 dataset [15] has been presented for single object tracking from 123 aerial videos. Videos in these datasets have been annotated for the ground truth object bounding boxes.
- **Gesture recognition datasets:** A gesture dataset recorded from a low-altitude hovering drone was presented in [17]. The UAV-Gesture dataset contains outdoor recorded of 13 drone commanding signals. NATOPS [42] is a similar indoor recorded gesture dataset consisting of 24 aircraft signal gestures.
- **Human action recognition datasets:** A large-scale VIRAT dataset [13] contains 550 videos covering a range of realistic and controlled human actions. The dataset has been recorded from both static and moving cameras (called VIRAT ground and aerial datasets). There are 23 outdoor event types with 29 h of videos. A limitation of the VIRAT aerial dataset is its low resolution of 480×720 pixels restricts algorithms from retrieving rich activity information from relatively small human subjects. A 4k resolution Okutama-Action [16] video dataset was introduced to detect 12 concurrent actions by multiple subjects. The dataset was recorded in a baseball field using two drones. Abrupt camera movements are present in the dataset making it more challenging for action recognition. However, the 90 degree elevation angle of the camera creates severe self-occlusions and perspective distortions in the videos. Other notable datasets in this category

are UCFARG [14], UCF aerial action [43], and Mini-drone [44]. UCF aerial and UCF ARG datasets have been recorded from an R/C-controlled blimp and a helium balloon respectively. Both datasets have similar action classes. UCF ARG is a multi-view dataset with synchronized ground, rooftop, and aerial camera views of the actions. However, UCF aerial action is a single-view dataset. Mini-drone was developed to study various aspects and definitions of privacy in aerial videos. The dataset was captured in a car park covering illicit, suspicious, and normal behaviors.

Most state-of-the-art human action recognition systems are implemented in high-end or graphics processing unit (GPU) powered computers, and those algorithms cannot perform as intended on typical embedded platforms due to the slowing of gains in commodity processor performance despite strong market pull [45]. The platform dependency and the limitations of processing power in drones have been bottlenecks to the development of onboard machine learning and AI techniques for human action or gesture recognition. Notably, action recognition from a mobile platform is not an extensively studied research area. Some notable works related to aerial image analysis are limited to offline processing of recorded videos [9,46–49]. In order to utilize the drone's mobile and operational capabilities to meet the emerging demand for drone technology, more studies are needed. In particular, the aerial action recognition studies need more datasets to develop technologies tailored to search and rescue, surveillance, and situational awareness. We tried to collect the data against relatively clutter-free backgrounds that were nonetheless normal outdoor environments.

3. Preparing the Dataset

This section explains the data collection process, the selected action classes, and the variations in the collected data. The dataset is compared with recently published aerial and ground video datasets.

3.1. Data Collection

The data collection was done on an unsettled road located in the middle of a wheat field. The aerial platform was a 3DR SOLO rotorcraft. The flight was slow and done at low altitudes (8–12 m). Videos were recorded using a GoPro Hero 4 Black camera with an anti-fish eye replacement lens (5.4 mm, 10 MP, IR CUT) and a 3-axis Solo gimbal. All videos in the dataset are HD (1920 × 1080) format captured at 25 fps.

A total of 13 actions were recorded while the drone was in hovering and following modes. In all videos, the subject was roughly in the middle of the frame and performed each action five to ten times except for *walking*, *jogging*, and *running*. For *walking*, *jogging*, and *running* actions, the subject was asked to traverse between two markers while the drone followed the subject from different directions.

When recording the actions, sometimes the drone drifted from its initial hovering position due to wind gusts. This adds random and characteristic camera motion to the videos corresponding to practical scenarios.

The total number of human subjects in the dataset was 10. The participants were asked to perform the actions in a selected section of the road. We divided the actions into the following three categories based on the viewpoint and the camera movement: *following*, *side-view*, and *front-view*. A detailed illustration of these three settings is provided in Figure 1.

Following actions: For the *walking*, *jogging*, and *running* actions, the subjects were recorded from four viewpoints. Each action was repeated four times while recording from the drone positioned at four different viewpoints in succession. As shown in Figure 1a, the left and right side views were recorded by maintaining a roughly perpendicular view with respect to the walking direction. However, due to mismatches in the moving speeds of the subject and the drone, in some videos, the subject was not always in the middle of the frame. Irrespective of the horizontal and vertical movements of the subject with respect to the camera frame, the videos were trimmed to include only the full body details of the person undertaking the particular action. We tried to follow and record the subject throughout the action, but in some videos, the drone was in hover mode or hover and follow modes combined.

In particular, when the subject was running fast, the recordings were mostly done in the hovering mode or had a little movement towards the direction in which the subject was moving. In terms of the camera motion, the recorded actions are from both moving and hovering modes.

Side-view actions: Five actions were recorded from both left and right sides of the subject when the drone was in hover mode; namely, *punching*, *hitting with a bottle*, *hitting with a stick*, *kicking*, and *stabbing*. We used mock foam weapons to represent sticks, bottles, and knives. Each action was repeated five to ten times (see Figure 1b). The actions were recorded with diversity across the class. For example, for the *kicking* action, each subject performed kicks towards the left and right sides of the camera frame. Each instance (left or right) had five to ten randomly performed kicks.

Front-view actions: *Clapping* and *waving hands* actions were recorded while the drone was hovering in front of the subject (see Figure 1c). In these videos, the subject was roughly in the middle of the frame and repeated each action five to ten times.

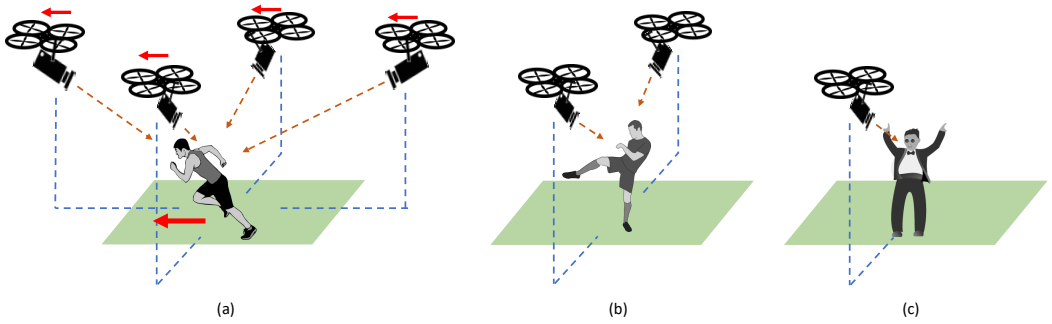

Figure 1. The actions were recorded in three different settings using a single drone. They are (**a**) following actions, (**b**) side-view actions, and (**c**) front-view actions.

3.2. Action Class Selection

The action categories described in Section 3.1 were selected considering three application scenarios.

(i) Analyzing human movements: This includes the *following actions* (*walking*, *jogging*, and *running*). Each action was recorded from four viewpoints from the front, back, and left and right sides of the subject. In the dataset, we consider the action recorded from the front and back views as one class and the action recorded from left and right sides as another class. For example, the *walking_side* action class has 20 videos recorded from left and right sides of 10 subjects, while *walking_front_back* has 20 videos recorded from front and back viewpoints. The 120 videos in these six action classes are also suitable for gait analysis and tracking.

(ii) Violence detection: The five *side-view actions* were selected from common violent acts. Punching and kicking are the most common bare hand actions of humans when fighting with each other [50]. It is also common to use some objects to hit or throw at each other [51]. To cover these scenarios, we included *hitting with a bottle*, *hitting with a stick*, and *stabbing*. Each action was recorded from left and right sides of the subject, capturing the maximum body surface area possible of these asymmetric actions. There are 100 videos showing the violent actions of 10 subjects.

(iii) Signaling the drone: In some scenarios, people might try to signal to the drone for help. The most popular way of signaling someone at a visible distance is using hand gestures. We selected two actions (*clapping* and *waving hands*) for this scenario and recorded from the front view as *front-view actions* (Figure 2). Each class contains 10 videos.

Figure 2. The action classes of the Drone-Action dataset. The images shown were selected randomly and cropped around the subject for clear demonstration.

3.3. Variations in Data

The actors were asked to perform each action five to ten times, with most of the actions being performed ten times. Each actor performed the actions differently, as they wished, capturing a variety of natural responses. All actors were volunteers from our research group and consisted of seven males and three females.

There are rich variations in the recorded actions in terms of the phase, orientation, camera movement, and the body shapes of the actors. The camera movements were caused by wind gusts and by the movement of the drone. In some videos, the scale rapidly changes. An example of scale change is the *running* action recorded from front or back viewpoints. The drone could not always maintain the speed of the subject in low-altitude flight. This resulted in scale variation of subjects during the course of the video. In some videos, the skin color of the actor is close to the background color. These variations create a challenging dataset for action recognition, and also makes it more representative of real-world situations.

3.4. Dataset Annotations

All of the videos have been annotated with subject ID, action class, and bounding box. Subject IDs are S1 to S10 for the ten actors. Along with bounding box annotations, body joint estimations computed using the widely used pose estimator OpenPose [52] were also included.

The experiments conducted here used three randomly generated split sets. These split sets were generated using the subject IDs associated with the videos; i.e., by avoiding overlapping of subject IDs between the train and test sets. Subject IDs can be used to generate customized split sets. The split sets used in our experiments are available for download with the dataset.

3.5. Dataset Summary

The dataset contains a total of 240 clips with 66919 frames. The number of actors in the dataset is 10 and they performed each action 5–10 times. All the videos are provided with 1920×1080 resolution and 25 fps. The average duration of each action was 11.15 sec. A summary of the dataset is given in Table 1. The total clip length and mean clip length of each class are represented on the left side (blue) and right side (amber) bar graphs of Figure 3 respectively. In Table 2, we compare our dataset with eight recently published video datasets. These datasets have helped progress research in action recognition, gesture recognition, event recognition, and object tracking.

Table 1. A summary of the dataset.

Feature	Value
# Actions	13
# Actors	10
# Clips	240
# Clips per class	10-20
Repetitions per class	5-10
Mean clip length	11.15 sec
Total duration	44.6 mins
# Frames	66919
Frame rate	25 fps
Resolution	1920 × 1080
Camera motion	Yes (hover and follow)
Annotation	Bounding box

Table 2. Comparison of recently published video datasets.

Dataset	Scenario	Purpose	Environment	Frames	Classes	Resolution	Year
UT Interaction [53]	Surveillance	Action recognition	Outdoor	36k	6	360 × 240	2010
NATOPS [42]	Aircraft signaling	Gesture recognition	Indoor	N/A	24	320 × 240	2011
VIRAT [13]	Drone, surveillance	Event recognition	Outdoor	Many	23	Varying	2011
UCF101 [33]	YouTube	Action recognition	Varying	558k	24	320 × 240	2012
J-HMDB [30]	Movies, YouTube	Action recognition	Varying	32k	21	320 × 240	2013
Mini-drone [44]	Drone	Privacy protection	Outdoor	23.3k	3	1920 × 1080	2015
Campus [54]	Surveillance	Object tracking	Outdoor	11.2k	1	1414 × 2019	2016
Okutama-Action [16]	Drone	Action recognition	Outdoor	70k	13	3840 × 2160	2017
UAV-Gesture [17]	Drone	Gesture recognition	Outdoor	37.2k	13	1920 × 1080	2018
Drone-Action	Drone	Action recognition	Outdoor	66.9k	13	1920 × 1080	2019

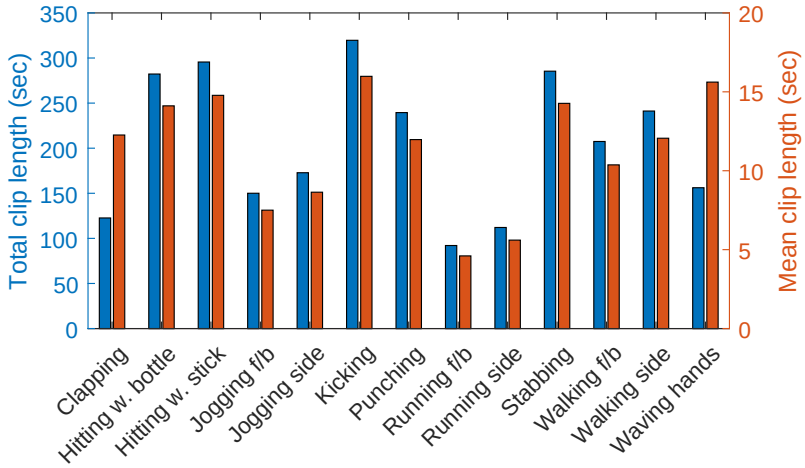

Figure 3. The total clip length (blue) and the mean clip length (amber) are shown in the same graph in seconds.

4. Experimental Results

We experimented with two popular feature types used in human action recognition; namely, pose-based Cheron et al.'s CNN features [29] and Jhuang et al.'s high-level pose features [30]. In this section, we report and compare experimental results using these two approaches.

4.1. High-Level Pose Features (HLPF)

HLPFs are formed by combining spatial and temporal properties of body keypoints throughout the action. We used the publicly available HLPF code [30] with minor modifications. HLPF was calculated using 15 keypoints (head, neck, shoulders, elbows, wrists, abdomen, hips, knees, and ankles). We used the pose estimator OpenPose [52] to find keypoints (see Figure 4a,b). In each frame, four spatial properties were calculated as follows:

- *Normalized positions*: Each key point was normalized with respect to the *belly* key point. A total of 30 descriptors were obtained (x and y coordinates of 15 joints).
- *Distance relationships*: The distance between two key points was calculated for all 15 key points, resulting in 105 descriptors.
- *Angular relationships*: A total of 1365 angles were calculated. Each angle was that of two vectors connecting a triplet of key points.
- *Orientation relationships*: 104 angles were calculated between each vector connecting two key points and the vector from *neck* to *abdomen*.

The temporal features were obtained by combining the following trajectories:

- *Cartesian trajectory*: 30 descriptors were obtained from the translation of 15 key points along x and y axes.
- *Radial trajectory*: The radial displacement of 15 key points provides 15 descriptors.
- *Distance relationship trajectory*: 105 descriptors were calculated from the differences between distance relationships.
- *Angle relationship trajectory*: 1365 descriptors were calculated from the differences between angle relationships.
- *Orientation relationship trajectory*: 104 descriptors were calculated from the differences between distance relationships.

An HLPF vector (including 3223 descriptor types) was generated by combining the above spatial and temporal descriptors. A codebook was generated by performing k-means clustering on each

descriptor type. The codeword size was selected to have $k = 20$. Classification was done using a SVM with linear kernels.

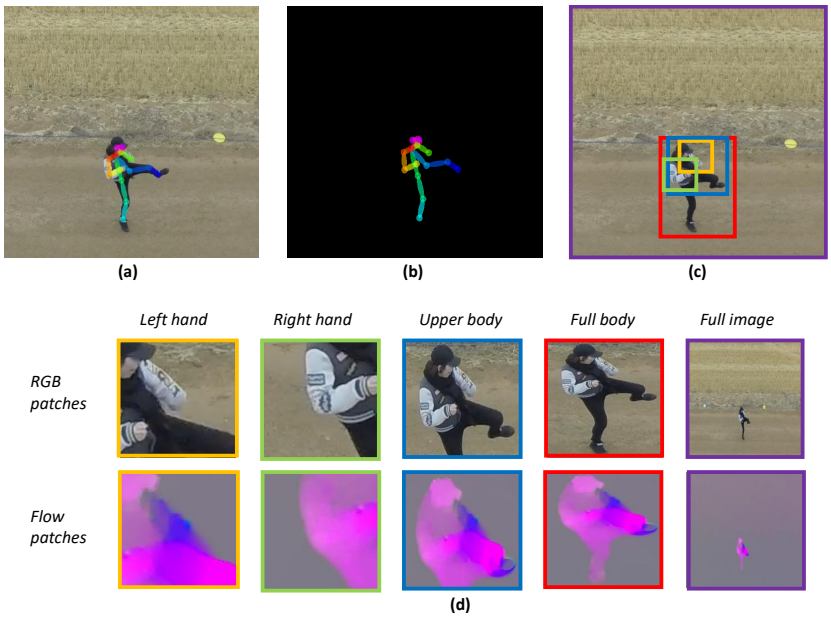

Figure 4. Extracted part patches (left hand, right hand, upper body, and full body) are shown using an image from the *kicking* class. (**a**) The estimated pose is shown with the background. The image is cropped around the subject for clear demonstration. (**b**) The keypoint locations are used compute high-level pose features (HLPF) and posed-based convolutional neural network (cP-CNN) features. (**c**) Part locations were determined based-on the keypoints. Purple color bounding box covers the full image. (**d**) RGB and flow patches were extracted from the RGB images and optical flow images respectively, using the bounding boxes shown in (**c**).

4.2. Pose-Based CNN (P-CNN)

P-CNN features [29] were created from person-centric motion and appearance features extracted using body joint positions. The P-CNN workflow is similar to Simonyan and Zisserman's two-stream CNN architecture [55]. For this work, we used the publicly available P-CNN code with minor modifications. The process of P-CNN feature extraction is shown in Figure 5. The main modules of the process are explained below.

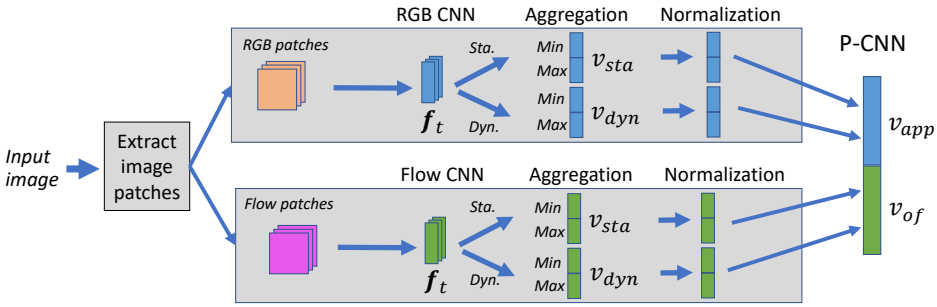

Figure 5. The steps involved in a P-CNN feature descriptor calculation. For a given image patch, one RGB and one flow CNN descriptor f_t is calculated per frame t.

4.2.1. Pose Estimation

We used the latest version of OpenPose [52] for pose estimation (see Figure 4a,b). OpenPose is a state-of-the-art method of pose estimation [56]. The pose estimator provides 18 key point estimations (eyes, ears, nose shoulders, elbows, wrists, hips, knees, and ankles). Head position was calculated by

taking the center of the eyes, ears, and nose key points. Sometimes, all of the key points in the head were not visible (e.g., from a side view). The *abdomen* key point was calculated by taking the centroid between the shoulder and hip key points. A total of 15 key points were extracted for the experiments (head, neck, shoulders, elbows, wrists, abdomen, hips, knees, and ankles).

4.2.2. Optical Flow Calculation

Optical flow was calculated between all consecutive frames of the video using Brox et al.'s implementation [57]. The calculated optical flow image was transformed to a heat map at the pixel level. The velocities along x and y axes, namely, v_x, v_y, were mapped to the range $[0, 255]$ by $[\hat{v}_x, \hat{v}_y] = s[v_x, v_y] + c$, with values outside the range $[0, 255]$ truncated at 0 or 255, according to [58]. The scalar s was calculated as $s = c/a$, where $a = 8$ was the maximum absolute value of the flow, and $c = 128$ was the new center of motion velocities.

4.2.3. Extracting Part Patches

The scale of each pose was used to extract consistent sized patches from the hands, upper body, and entire body of the subject. The scale for each pose was obtained by dividing the pose height (in pixels) by 240. In the image coordinate system, height was calculated using the maximum difference in y-coordinates between the head and ankles.

A region of interest around a body part was extracted by adding some margin to the top-left and bottom-right key point positions of the body part—we called this a *part patch*. The margin was calculated as $C \times S$, where C was selected to be 100 pixels, and S was the scale. The extracted patch was resized to 224×224 to match the input size of the CNN model. Figure 4 illustrates the process of extracting part patches using an image from the *Kicking* class.

4.2.4. CNN Feature Aggregation

For each body part and full image, the appearance (RGB) and optical flow patches were extracted, and their CNN features were computed using two pre-trained networks (see Figure 5). For appearance patches, the publicly available "VGG-f" network [59] was used, whereas for optical flow patches, the motion network from Gkioxari and Malik's Action Tube implementation [58] was used. Both networks have the same architecture with five convolutional and three fully-connected layers. For each body part and frame t, a frame descriptor f_t was obtained from the output of the second fully-connected layer. Static and dynamic features were separately aggregated over time to obtain a static video descriptor v_{sta} and a dynamic video descriptor v_{dyn} respectively.

Min and *max* aggregation schemes were used to calculate the minimum and maximum values of each descriptor dimension k ($k \in \{1, ..., 4096\}$) over T frames.

$$\begin{aligned} m_k &= \min_{1 \le t \le T} f_t(k), \\ M_k &= \max_{1 \le t \le T} f_t(k). \end{aligned} \tag{1}$$

For static and dynamic video descriptors, the time-aggregated video descriptors were concatenated as follows:

$$v_{sta} = [m_1, \ldots, m_k, M_1, \ldots, M_k]^\top, \tag{2}$$

$$v_{dyn} = [\Delta m_1, \ldots, \Delta m_k, \Delta M_1, \ldots, \Delta M_k]^\top. \tag{3}$$

Δ represents temporal differences in the video descriptors. The aggregated features (v_{sta} and v_{dyn}) were normalized and concatenated over the number of body parts to obtain appearance features v_{app} and flow features v_{of}. The final P-CNN descriptor was obtained by concatenating v_{app} and v_{of}.

4.3. Performance Evaluation

The evaluation metric selected for the experiment was accuracy. Accuracy was calculated using the scores returned by the action classifiers. We experimented with the dataset by generating three split sets. Each split set was generated randomly with a 70 : 30 train-to-test ratio using the subject ID.

We report a mean classification accuracy of 64.36% for HLPF. As explained in Section 4.1, classification was performed using linear kernels. Accuracies for the three split sets were 63.89%, 68.09%, and 61.11%.

The mean accuracy reported for P-CNN was 75.92%. The accuracies of the three split set were 72.22%, 81.94%, and 73.61%. Here, the classification was performed using a linear SVM.

In Table 3, we compare our dataset with two other action recognition datasets; namely, JHMDB and MPII. Both of these datasets used Cherian et al.'s pose estimation algorithm [60]. The performance metrics for the JHMDB and MPII datasets were taken from [29].

Two confusion matrices were calculated (see Figure 6) for assessing the action recognition accuracies of the HLPF and P-CNN approaches. Both HLPF and P-CNN had some difficulty distinguishing between *hitting with bottle* and *stabbing* actions, which were associated with the lowest scores in the confusion matrix for P-CNN and noticeably low scores for HLPF. The main reason for this was the similar patterns of body joint movements for these two classes of actions, although HLPF does not take into account the characteristics of associated objects. Another possibility specific to P-CNN was that the color of the bottle used in *hitting with bottle* was close to the background color and sometimes difficult to identify. The classifiers can also confuse between *jogging* and *running* actions, as the speeds of these actions varied from one actor to another. Overall, P-CNN showed better performance for most of the classes.

As an alternative measure of accuracy, Cohen's kappa coefficient [61] was calculated for both the HLPF and PCNN confusion matrices. Cohen's kappa was originally proposed in the area of psychology as a measure of agreement between two "judges" that each classifies some objects into some mutually exclusive categories. The coefficient was defined as $\kappa \triangleq \frac{p_o - p_e}{1 - p_e}$, where p_o is the proportion of agreements, and p_e is the proportion of chance agreements. According to Landis and Koch [62], a κ value of < 0 should be interpreted as no agreement, 0–0.20 as slight, 0.21–0.40 as fair, 0.41–0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1 as almost perfect agreement. In the context of multi-class classification, an agreement is a true positive. Given confusion matrix $[m_{i,j}]$ and k classes,

$$p_o = \frac{m_{1,1} + m_{2,2} + \cdots + m_{k,k}}{\sum_{i,j \in \{1,\ldots,k\}} m_{i,j}}, \qquad p_e = \frac{\sum_{l=1}^{k} (\sum_{j=1}^{k} m_{l,j} \cdot \sum_{i=1}^{k} m_{i,l})}{\sum_{i,j \in \{1,\ldots,k\}} m_{i,j}}.$$

Using Cardillo's code [63], Cohen's κ coefficients for the HLPF and P-CNN confusion matrices were found to be 0.6433 and 0.7593 respectively. As per Landis and Koch's [62] interpretation, both κ coefficients indicated substantial agreement, with the κ for P-CNN being conspicuously higher than that for HLPF, and in fact nearly falling in the region of almost perfect agreement.

We also performed experiments by (i) combining P-CNN and HLPF features to form a new feature vector, and (ii) fusing the classification scores of P-CNN and HLPF classifiers. In both cases, the performance was degraded below the accuracy reported for P-CNN alone. Therefore, we report the P-CNN feature-based action classification as the baseline algorithm used for our dataset.

Table 3. The best reported HLPF and P-CNN action recognition results for different datasets.

Dataset	Remarks	HLPF	P-CNN
JHMDB	Res: 320 × 240, Pose [60]	25.30	61.10
MPII Cooking	Res: 1624 × 1224, Pose [60]	32.60	62.30
Drone-Action	Res: 1920 × 1080, OpenPose [52]	64.36	75.92

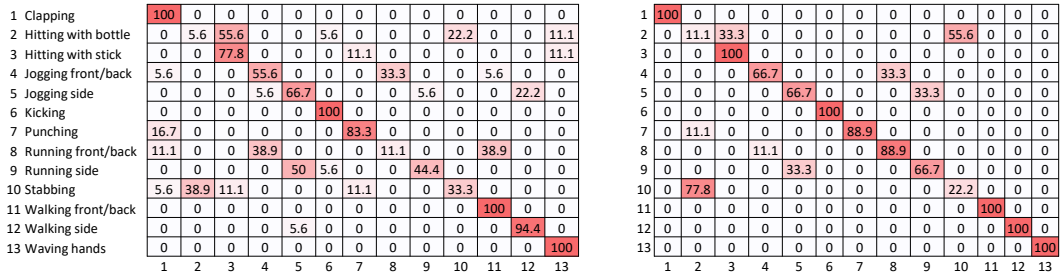

Figure 6. Confusion matrices calculated as percentage accuracies for HLPF (**on the left**) and P-CNN (**on the right**).

5. Discussion

HLPF computation is based only on the spatial and temporal relationships among the key points. It does not consider any additional information associated with the actions, such as the objects used in the actions. In comparison, P-CNN computation includes the body parts and action-specific objects visible in the part patches. This results in a higher action recognition accuracy for P-CNN compared to HLPF. This observation is applicable to all three datasets used in this work.

In both HLPF and P-CNN approaches, actions with similar body joint movements can be confused as one another, resulting in an overall reduction in accuracy. Both feature descriptors performed well with distinctive actions. It was noted that HLPF accuracy decreased when the action was associated with additional objects, while P-CNN was robust whether additional objects were involved or not.

We used the state-of-the-art pose estimator OpenPose to find key points. However, as with all estimators, sometimes it missed the key points due to varying human scales and viewpoint occlusions. The estimator can be made more robust to scale variation by increasing the input image size of the pose estimator model, at the expense of computational overhead. Perspective distortion caused by the use of an aerial camera is minimal in most of the videos, although in the front/back *following* actions, the perspective distortion varies when the subject moves closer to or further away from the drone. This distortion can be compensated for to some extent using an approach similar to that reported by Perera et al. [64].

This dataset contains only single-person actions. As explained in Sections 1 and 2, drone-based human action recognition is still an emerging research area. The closest research area to this is gesture-based drone control. We believe that single-person actions provide an opportunity for many researchers to start or advance their research in aerial action recognition. That is why this dataset focuses on the first step in aerial action recognition: single-person action recognition. Our ongoing research includes extending this dataset to a multi-person version.

In multi-person action recognition studies, normally, the actions are broadly categorized based on the event. A relevant dataset and approach are described in the VIRAT aerial dataset [13]. When detecting the actions of a group of people, each individual in the scene should be detected and analyzed for the collective or individual actions [16]. However, depending on the scene, multiple subjects can improve the detection accuracy in the actions (e.g., fighting).

This dataset has been created to help progress research in aerial action recognition. The work presented in this paper is intended to provide a sufficient amount of data to cover some common human actions. These actions can be visible from a low-altitude slow-flying drone. When recording videos, we used a relatively clutter-free background and maximized the image details by keeping the subject roughly in the middle of the frame throughout the drone's hover and follow modes. Therefore, our dataset does not include videos with complex backgrounds and challenging flight maneuvers.

6. Conclusions

We presented an action recognition dataset, called Drone-Action, recorded by a hovering drone. The dataset contains 240 HD videos lasting for a total of 44.6 minutes. The dataset was prepared

using 13 selected actions from common outdoor human actions. The actions were recorded from 10 participants in an outdoor setting while the drone was in hover or follow mode. The rich variation of body size, camera motion, and phase, makes our dataset challenging for action recognition. The dataset contains videos recorded from a low-altitude, slow-flying drone. The action classes, the relatively large images of human subjects and high resolution image details extend the dataset's applicability to a wider research community.

We evaluated this new dataset using HLPF and P-CNN descriptors, and reported an overall baseline action recognition accuracy of 75.92% using P-CNN. This dataset is useful for research involving surveillance, situational awareness, general action recognition, violence detection, and gait recognition. The Drone-Action dataset is available at https://asankagp.github.io/droneaction.

Author Contributions: The original draft preparation was done by A.G.P. Both Y.W.L.'s and J.C. contributions were to aid in conceptualization, reviewing, and editing.

Funding: This project was partly supported by Project Tyche, the Trusted Autonomy Initiative of the Defence Science and Technology Group (grant number myIP6780).

Acknowledgments: We thank the student volunteers who participated in the data collection.

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

References

1. Gonçalves, J.; Henriques, R. UAV photogrammetry for topographic monitoring of coastal areas. *ISPRS J. Photogramm. Remote Sens.* **2015**, *104*, 101–111. [CrossRef]
2. Barbedo, J.G.A. A Review on the Use of Unmanned Aerial Vehicles and Imaging Sensors for Monitoring and Assessing Plant Stresses. *Drones* **2019**, *3*, 40. [CrossRef]
3. Al-Kaff, A.; Moreno, F.M.; José, L.J.S.; García, F.; Martín, D.; de la Escalera, A.; Nieva, A.; Garcéa, J.L.M. VBII-UAV: Vision-Based Infrastructure Inspection-UAV. In *Recent Advances in Information Systems and Technologies*; Rocha, Á., Correia, A.M., Adeli, H., Reis, L.P., Costanzo, S., Eds.; Springer International Publishing: Cham, Switzerland, 2017; pp. 221–231.
4. Erdelj, M.; Natalizio, E.; Chowdhury, K.R.; Akyildiz, I.F. Help from the Sky: Leveraging UAVs for Disaster Management. *IEEE Pervasive Comput.* **2017**, *16*, 24–32. [CrossRef]
5. Peschel, J.M.; Murphy, R.R. On the Human–Machine Interaction of Unmanned Aerial System Mission Specialists. *IEEE Trans. Hum.-Mach. Syst.* **2013**, *43*, 53–62. [CrossRef]
6. Chahl, J. Unmanned Aerial Systems (UAS) Research Opportunities. *Aerospace* **2015**, *2*, 189–202. [CrossRef]
7. Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018.
8. Razakarivony, S.; Jurie, F. Vehicle detection in aerial imagery : A small target detection benchmark. *J. Vis. Commun. Image Represent.* **2016**, *34*, 187–203. [CrossRef]
9. Krajewski, R.; Bock, J.; Kloeker, L.; Eckstein, L. The highD Dataset: A Drone Dataset of Naturalistic Vehicle Trajectories on German Highways for Validation of Highly Automated Driving Systems. In Proceedings of the 2018 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; pp. 2118–2125. [CrossRef]
10. Interstate 80 Freeway Dataset. 2019. Available online: :https://www.fhwa.dot.gov/publications/research/operations/06137/index.cfm (accessed on 2 November 2019).
11. Zhu, P.; Wen, L.; Bian, X.; Haibin, L.; Hu, Q. Vision Meets Drones: A Challenge. *arXiv* **2018**, arXiv:1804.07437.
12. Carletti, V.; Greco, A.; Saggese, A.; Vento, M. Multi-Object Tracking by Flying Cameras Based on a Forward-Backward Interaction. *IEEE Access* **2018**, *6*, 43905–43919. [CrossRef]
13. Oh, S.; Hoogs, A.; Perera, A.; Cuntoor, N.; Chen, C.C.; Lee, J.T.; Mukherjee, S.; Aggarwal, J.K.; Lee, H.; Davis, L.; et al. A large-scale benchmark dataset for event recognition in surveillance video. In Proceedings of the CVPR 2011, Providence, RI, USA, 20–25 June 2011; pp. 3153–3160. [CrossRef]
14. University of Central Florida. UCF-ARG Data Set. 2011. Available online: http://crcv.ucf.edu/data/UCF-ARG.php (accessed on 2 November 2019).

15. Mueller, M.; Smith, N.; Ghanem, B. A Benchmark and Simulator for UAV Tracking. In Proceedings of the Computer Vision—ECCV 2016, Amsterdam, The Netherlands, 8–16 October 2016; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 445–461.
16. 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 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017; pp. 2153–2160. [CrossRef]
17. 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) Workshops, Munich, Germany, 8–14 September 2018; pp. 117–128. [CrossRef]
18. Natarajan, K.; Nguyen, T.D.; Mete, M. Hand Gesture Controlled Drones: An Open Source Library. In Proceedings of the 2018 1st International Conference on Data Intelligence and Security (ICDIS), South Padre Island, TX, USA, 8–10 April 2018; pp. 168–175. [CrossRef]
19. Lee, J.; Tan, H.; Crandall, D.; Šabanović, S. Forecasting Hand Gestures for Human-Drone Interaction. In Proceedings of the Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, Chicago, IL, USA, 5–8 March 2018; pp. 167–168. [CrossRef]
20. Hsu, H.J.; Chen, K.T. DroneFace: An Open Dataset for Drone Research. In Proceedings of the 8th ACM on Multimedia Systems Conference, Taipei, Taiwan, 20–23 June 2017; pp. 187–192. [CrossRef]
21. Kalra, I.; Singh, M.; Nagpal, S.; Singh, R.; Vatsa, M.; Sujit, P.B. DroneSURF: Benchmark Dataset for Drone-based Face Recognition. In Proceedings of the 2019 14th IEEE International Conference on Automatic Face Gesture Recognition (FG 2019), Lille, France, 14–18 May 2019; pp. 1–7. [CrossRef]
22. Carletti, V.; Greco, A.; Saggese, A.; Vento, M. An intelligent flying system for automatic detection of faults in photovoltaic plants. *J. Ambient Intell. Hum. Comput.* **2019**. [CrossRef]
23. Avola, D.; Cinque, L.; Foresti, G.L.; Martinel, N.; Pannone, D.; Piciarelli, C. A UAV Video Dataset for Mosaicking and Change Detection From Low-Altitude Flights. *IEEE Trans. Syst. Man Cybern. Syst.* **2018**. [CrossRef]
24. Sensefly Mosaic Datasets. 2019. Available online: https://www.sensefiy.com/drones/example-datasets.html (accessed on 2 November 2019).
25. Lottes, P.; Khanna, R.; Pfeifer, J.; Siegwart, R.; Stachniss, C. UAV-based crop and weed classification for smart farming. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 3024–3031. [CrossRef]
26. Monteiro, A.; von Wangenheim, A. Orthomosaic Dataset of RGB Aerial Images for Weed Mapping. 2019. Available online: http://www.lapix.ufsc.br/weed-mapping-sugar-cane (accessed on 2 November 2019).
27. Herath, S.; Harandi, M.; Porikli, F. Going deeper into action recognition: A survey. *Image Vis. Comput.* **2017**, *60*, 4–21. [CrossRef]
28. Mabrouk, A.B.; Zagrouba, E. Abnormal behavior recognition for intelligent video surveillance systems: A review. *Expert Syst. Appl.* **2018**, *91*, 480–491. [CrossRef]
29. Cheron, G.; Laptev, I.; Schmid, C. P-CNN: Pose-Based CNN Features for Action Recognition. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
30. Jhuang, H.; Gall, J.; Zuffi, S.; Schmid, C.; Black, M.J. Towards Understanding Action Recognition. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3192–3199. [CrossRef]
31. Schuldt, C.; Laptev, I.; Caputo, B. Recognizing human actions: A local SVM approach. In Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, UK, 26 August 2004; Volume 3, pp. 32–36.
32. Blank, M.; Gorelick, L.; Shechtman, E.; Irani, M.; Basri, R. Actions as space-time shapes. In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05), Beijing, China, 17–21 October 2005; Volume 2, pp. 1395–1402. [CrossRef]
33. Soomro, K.; Zamir, A.R.; Shah, M. *UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild*; Technical Report; UCF Center for Research in Computer Vision: Orlando, FL, USA, 2012.
34. Zhang, W.; Zhu, M.; Derpanis, K.G. From Actemes to Action: A Strongly-Supervised Representation for Detailed Action Understanding. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 2248–2255. [CrossRef]

35. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-Scale Video Classification with Convolutional Neural Networks. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 24–27 June 2014; pp. 1725–1732. [CrossRef]
36. Heilbron, F.C.; Escorcia, V.; Ghanem, B.; Niebles, J.C. ActivityNet: A large-scale video benchmark for human activity understanding. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 961–970. [CrossRef]
37. Feichtenhofer, C.; Pinz, A.; Zisserman, A. Convolutional Two-Stream Network Fusion for Video Action Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1933–1941. [CrossRef]
38. Abu-El-Haija, S.; Kothari, N.; Lee, J.; Natsev, A.; Toderici, G.; Varadarajan, B.; Vijayanarasimhan, S. YouTube-8M: A Large-Scale Video Classification Benchmark. *arXiv* **2016**, arXiv:1609.08675.
39. Kay, W.; Carreira, J.; Simonyan, K.; Zhang, B.; Hillier, C.; Vijayanarasimhan, S.; Viola, F.; Green, T.; Back, T.; Natsev, A.; et al. The Kinetics Human Action Video Dataset. *arXiv* **2017**, arXiv:1705.06950.
40. Zhao, H.; Yan, Z.; Torresani, L.; Torralba, A. HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization. *arXiv* **2019**, arXiv:1712.09374.
41. Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; Tian, Q. The Unmanned Aerial Vehicle Benchmark: Object Detection and Tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
42. Song, Y.; Demirdjian, D.; Davis, R. Tracking body and hands for gesture recognition: NATOPS aircraft handling signals database. *Face Gesture* **2011**, 500–506. [CrossRef]
43. University of Central Florida. UCF Aerial Action Dataset. 2011. Available online: http://crcv.ucf.edu/data/UCF_Aerial_Action.php (accessed on 02 Novomber 2019).
44. Bonetto, M.; Korshunov, P.; Ramponi, G.; Ebrahimi, T. Privacy in mini-drone based video surveillance. In Proceedings of the 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Ljubljana, Slovenia, 4–8 May 2015; Volume 4, pp. 1–6. [CrossRef]
45. Ovtcharov, K.; Ruwase, O.; Kim, J.Y.; Fowers, J.; Strauss, K.; Chung, E.S. Accelerating deep convolutional neural networks using specialized hardware. *Microsoft Res. Whitepaper* **2015**, *2*, 1–4.
46. Rudol, P.; Doherty, P. Human Body Detection and Geolocalization for UAV Search and Rescue Missions Using Color and Thermal Imagery. In Proceedings of the 2008 IEEE Aerospace Conference, Big Sky, MT, USA, 1–8 March 2008; pp. 1–8. [CrossRef]
47. Oreifej, O.; Mehran, R.; Shah, M. Human identity recognition in aerial images. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 709–716. [CrossRef]
48. Yeh, M.C.; Chiu, H.K.; Wang, J.S. Fast medium-scale multiperson identification in aerial videos. *Multimed. Tools Appl.* **2016**, *75*, 16117–16133. [CrossRef]
49. Al-Naji, A.; Perera, A.G.; Chahl, J. Remote monitoring of cardiorespiratory signals from a hovering unmanned aerial vehicle. *BioMedical Eng. OnLine* **2017**, *16*, 101. [CrossRef]
50. De Souza, F.D.; Chavez, G.C.; do Valle, E.A., Jr.; Araújo, A.D.A. Violence Detection in Video Using Spatio-Temporal Features. In Proceedings of the 2010 23rd SIBGRAPI Conference on Graphics, Patterns and Images, Gramado, Brazil, 30 August–3 September 2010; pp. 224–230. [CrossRef]
51. Datta, A.; Shah, M.; Lobo, N.D.V. Person-on-person violence detection in video data. In Proceedings of the Object Recognition Supported by User Interaction for Service Robots, Quebec City, QC, Canada, 11–15 August 2002. [CrossRef]
52. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. In Proceedings of the CVPR, Honolulu, HI, USA, 21–26 July 2017.
53. Ryoo, M.S.; Aggarwal, J.K. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 1593–1600. [CrossRef]
54. Robicquet, A.; Sadeghian, A.; Alahi, A.; Savarese, S. Learning Social Etiquette: Human Trajectory Understanding In Crowded Scenes. In Proceedings of the Computer Vision—ECCV, Amsterdam, The Netherlands, 11–14 October 2016; Leibe, B., Matas, J., Sebe, N., Welling, M., Eds.; Springer International Publishing: Cham, Switzerland, 2016; pp. 549–565.

55. Simonyan, K.; Zisserman, A. Two-Stream Convolutional Networks for Action Recognition in Videos. In Proceedings of the 27th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 568–576.
56. Zhao, M.; Li, T.; Alsheikh, M.A.; Tian, Y.; Zhao, H.; Torralba, A.; Katabi, D. Through-Wall Human Pose Estimation Using Radio Signals. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18-22 June 2018; pp. 7356–7365. [CrossRef]
57. Brox, T.; Bruhn, A.; Papenberg, N.; Weickert, J. High Accuracy Optical Flow Estimation Based on a Theory for Warping. In *Proceedings of the Computer Vision—ECCV 2004*; Pajdla, T., Matas, J., Eds.; Springer: Berlin/Heidelberg, Germany, 2004; pp. 25–36.
58. Gkioxari, G.; Malik, J. Finding Action Tubes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
59. Chatfield, K.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Return of the Devil in the Details: Delving Deep into Convolutional Nets. *arXiv* **2014**, arXiv:1405.3531v4.
60. Cherian, A.; Mairal, J.; Alahari, K.; Schmid, C. Mixing Body-Part Sequences for Human Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 24–27 June 2014.
61. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [CrossRef]
62. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. [CrossRef] [PubMed]
63. Cardillo, G. Compute the Cohen's Kappa (Version 2.0.0.0). 2018. Available online: http://www.mathworks.com/matlabcentral/fileexchange/15365 (accessed on 2 November 2019).
64. Perera, A.G.; Law, Y.W.; Chahl, J. Human Pose and Path Estimation from Aerial Video Using Dynamic Classifier Selection. *Cogn. Comput.* **2018**, *10*, 1019–1041. [CrossRef]