*remote sensing*

MDPI

*Article*

# Detection and Localisation of Life Signs from the Air Using Image Registration and Spatio-Temporal Filtering

**Asanka G. Perera [1,*], Fatema-Tuz-Zohra Khanam [1], Ali Al-Naji [1,2] and Javaan Chahl [1,3]**

[1] School of Engineering, University of South Australia, Mawson Lakes, SA 5095, Australia; fatema-tuz-zohra.khanam@mymail.unisa.edu.au (F.-T.-Z.K.); ali_al_naji@mtu.edu.iq (A.A.-N.); Javaan.Chahl@unisa.edu.au (J.C.)

[2] Electrical Engineering Technical College, Middle Technical University, Baghdad 1022, Iraq

[3] Joint and Operations Analysis Division, Defence Science and Technology Group, Melbourne, VIC 3207, Australia

* Correspondence: asanka.perera@mymail.unisa.edu.au

Received: 14 December 2019; Accepted: 4 February 2020 ; Published: 9 February 2020


**Abstract:** In search and rescue operations, it is crucial to rapidly identify those people who are alive from those who are not. If this information is known, emergency teams can prioritize their operations to save more lives. However, in some natural disasters the people may be lying on the ground covered with dust, debris, or ashes making them difficult to detect by video analysis that is tuned to human shapes. We present a novel method to estimate the locations of people from aerial video using image and signal processing designed to detect breathing movements. We have shown that this method can successfully detect clearly visible people and people who are fully occluded by debris. First, the aerial videos were stabilized using the key points of adjacent image frames. Next, the stabilized video was decomposed into tile videos and the temporal frequency bands of interest were motion magnified while the other frequencies were suppressed. Image differencing and temporal filtering were performed on each tile video to detect potential breathing signals. Finally, the detected frequencies were remapped to the image frame creating a life signs map that indicates possible human locations. The proposed method was validated with both aerial and ground recorded videos in a controlled environment. Based on the dataset, the results showed good reliability for aerial videos and no errors for ground recorded videos where the average precision measures for aerial videos and ground recorded videos were 0.913 and 1 respectively.

**Keywords:** search and rescue; drone; breathing detection; human detection

## 1. Introduction

Search and rescue (SAR) operations play a vital role in a any disaster stricken area. Usually such areas are difficult to reach due to impassable roads, flooded waterways and unsafe air infrastructure. Rescue services include searching for missing people and treating the injured at disaster sites. In SAR operations, time is a critical factor as exposure, dehydration, radiation dose and haemorrhaging continue [1]. The specific location of the casualties and even their number is usually unknown. Generally the area affected is large and it must be searched as thoroughly as possible within the shortest possible period. Ground search for victims is usually difficult for rescue workers due to uneven surfaces, poor weather and difficult terrain. Contamination may also occur in the same area as a result of nuclear disasters or biohazards. Extreme temperatures, flash flooding and fires might make the environment more hazardous for dismounted rescuers moving than for stationary casualties. SAR can be a hazardous task causing mental and physical fatigue to rescuers [2].

Ground robots can be used instead of human rescuers for search and rescue purposes [3–5]. However, they are expensive and need invasive equipment and need human operators for moving around the disaster zone. Because of limited mobility of wheeled and tracked platforms, it is sometimes more difficult for rescue robots to get into a disaster zone than for rescue workers and rescue dogs. To this day, ground robot mobility is not guaranteed one metre off a formed road outside urban areas.

Drones, also known as unmanned aerial vehicles (UAVs) can be an alternate solution to minimize these issues as drones have mobility and height advantages over humans and ground robots [6]. Firstly, drones can be sent to any place without having prior information about the actual conditions of the targeted area. After an initial survey at altitude it can be determined what altitudes and areas are available to the operation. This minimizes the risk to rescue workers of injury or death. Additionally, by using modern detection and tracking systems, drones can examine a large area in a short period by compiling high resolution aerial images or recording a video of the targeted location. RGB, near infrared and thermal cameras combined with machine learning (ML) or other techniques can be used to detect and track humans [7].

Detecting or tracking a human is not sufficient for search and rescue missions in disaster zones. It is also necessary to detect whether the human is alive or not, so that living persons can be rescued and treated first. Ideally judgements are made about the prospect of saving a person based on their vital signs and their immediate need to be extracted. Physiological parameters such as heart rate and breathing rate are the parameters that help to identify whether the person is alive or not. With the development of computer vision technology, remote monitoring of physiological parameters is becoming more popular which not only plays an important role in clinical sectors but also in nonclinical sectors including war zones and natural disaster zones [8,9].

Sometimes, the disaster stricken areas are covered with debris, dust or ash making the standard computer vision approaches less effective in search and rescue missions. When human shapes are not detectable due to texture and when body heat is not distinguishable from ground heat in a warm environment, the sensor readings of a SAR drone might have limited use. There is a need to scan the particular area and identify living humans despite challenging visual artefacts. In this study, we propose a novel method to scan such an area and detect living humans only by their breathing. We examined past studies related to drone-based search and rescue operations, specifically studies focused on visual sensor data for human detection and tracking. To our knowledge, this is the first study that proposes to scan a particular area from aerial videos to detect any living human lying on the ground by their breathing.

Our method works for humans in the open and additionally, since the solution does not need to detect the human form, we have shown that it can work for live humans covered with debris. We analyze the aerial images by stabilizing them at two levels. First, by stabilizing the raw image sequence and then, by stabilizing the decomposed tile videos. We used a sliding window to decompose the original stabilized video into tiles. Motion magnification was applied to amplify potential chest movements and to attenuate other frequencies. The estimated breathing rate of the tile video was mapped to a localization map to indicate potential human subjects. We validated our approach with ten aerial videos and four ground recorded videos. The ground recorded videos detected all living subjects while the aerial videos showed promising detection accuracy in the presence of few false positives caused by camera noise and motion.

The main contributions of this study are as follows:

- The standard video stabilization techniques transform the full image plane into a new plane based on the scene geometry or the image features. These techniques focus on a global transformation. When the video is recorded by a moving camera this global transformation cannot compensate for all local scene geometry changes. To address this issue, we propose a two-step stabilization method: (i) globally stabilize the video and (ii) decompose the globally stabilized the video into video tiles and stabilize individual tile videos.

- We propose a complete process for living human detection from drone-recorded videos. The process comprises of video stabilization, motion magnification and temporal filtering. We leverage the tools and methods available in the literature [10–13] and develop a process for this novel application. This is an extension to the work presented by Al-Naji et al. [11].
- We collected a dataset in a controlled environment and experimentally proved that a living humans can be detected purely by analysing their breathing motion using drone videos.

The rest of this paper is organized as follows. Section 2 discusses closely related work on vision-based search and rescue studies and improvements related to human detection and tracking. Section 3 describes our methodology. Section 4 reports experimental results. Discussion of issues and potential improvements are presented in Section 5. Section 6 concludes.

## 2. Related Work

The survey [14] compares studies that use drones in search and rescue operations published up to the year 2018. Here, we discuss some studies related to our proposed work. Several studies have used drones for search and rescues operations by incorporating different techniques. For example, Andriluka et al. [15] proposed a method by combining multiple models to automatically detect people on the ground from images captured by the on-board daylight camera on a quadrotor drone. Camara et al. [16] proposed a drone fleet to help rescuers in disaster zones. Solutions have been proposed to leverage the state-of-the-art human detection techniques. Sulistijono et al. [17] presented a study to detect disaster victims in highly cluttered scenarios. Lygouras et al. [18] also addressed a similar problem of detecting swimmers in disaster areas in the presence of complex backgrounds. Both studies proposed to use deep learning techniques for victim detection and drones for efficient data capture. Drone-based people detection was also proposed for use in avalanche search and rescue operations [19]. Al-Kaff et al. [20] developed an appearance-based tracking algorithm using the color and depth data obtained from a Kinect-V2 sensor on a drone. For aerial search and rescue purposes, they proposed two algorithms: (i) A multi-object tracking algorithm was proposed to detect victims lying on the ground in unconstrained poses and (ii) a semi-autonomous reactive control was proposed for performing safe and smooth approach maneuvers to the detected persons. Their system was efficient for detecting humans in many poses. However, it was limited to detecting human forms over short distances. A drone equipped with a voice recognition system was proposed by Yamazaki et al. [21] to detect victims in a disaster stricken area. Nevertheless, all of the studies discussed above were only interested in detecting human forms and their location but they did not consider whether the victim was alive or not, so no life detection was attempted in these studies.

Here, we discuss some important topics related to aerial search and rescue operations.

### *2.1. Using Thermal and IR Cameras for Human Detection*

Some works have used thermal sensors or multiple sensors for the purpose of detecting and tracking humans. Portmann et al. [22] proposed a detection and tracking algorithm based on a robust background subtraction method and a particle filter guided detector for detecting and tracking humans from aerial thermal images. A multiple sensor-based tracking approach was presented by Kang et al. [23]. They developed a method for detecting and tracking multiple people by considering electro-optical (EO) and infra-red (IR) sensors using a joint probability model. To detect humans, Doherty and Rudol [24] used both thermal and color cameras on an autonomous drone in outdoor environments. The thermal camera was used to prefilter promising image locations and subsequently a visual object detector was used to verify them. Rivera et al. [25] presented a human detection and geolocation system using aerial drones to search for victims in SAR operations. They designed modular equipment by combining thermal and colour cameras and a geolocation module attached to a drone. The best results of their experiments were achieved during night operations. By integrating visible and thermal images recorded by the drone, Blondel et al. [26] developed a fast, robust human detection

technique for SAR operations. The pitch and roll-trained detector (PRD) and an accurate search space reduction technique were used.

However, these studies have several limitations with low resolution, short-range detection and motion artefacts caused by camera movement. Additionally, the integration of multiple sensors for human detection is costly and adds an additional payload beyond the standard off the shelf arrangement. In previous studies related to SAR, life signs detection of humans were not considered. When there is a thermal difference between the human body and background, thermal cameras are likely to identify living persons. This makes thermal cameras see less contrast between living people and warm backgrounds. We used only an RGB camera for our study but we do note the practical importance of using a thermal camera in parallel to the RGB camera to enhance the implementation for real scenarios.

### *2.2. Real-Time Data Transmission*

The on-board processor of the drone has limited computational power and is not suitable to run computationally heavy software components. To address this issue, several video transmission techniques have been developed in recent literature. The main aims of such a system is to achieve a real-time data transmission link between the drone and the ground controller and to use a powerful computer on the ground for faster processing. Several strategies have been presented previously to use the existing cellular telephone network [27,28] to selectively transmit frames [29]. Video transmission is an important part of SAR image analysis. However, we limit the scope of our work to offline image processing to demonstrate the principle.

### *2.3. Video Stabilization*

Video stabilization is a necessary step in aerial video analysis. Most aerial videos suffer from uncommanded camera motions caused by platform movement and wind. Different video stabilization techniques have been proposed in the literature. Some notable studies have been reported using 2D feature trajectories [30], 3D camera motion [31] and motion in-painting and sharpened image pixels [32]. Key point-based video stabilization has shown promising results in aerial videos [33,34]. In these studies, the image sequence is stabilized over a reference frame. There is a sufficiently large distance between the aerial camera and scene geometry of interest. Therefore, most of the aerial video-related studies focus on key point distribution in the 2D image plane. Video stabilization is a computationally expensive operation. Liu et al. [30] proposed a fast stabilization algorithm for video in which accurate scene reconstruction is not necessary. However, achieving both accuracy and real-time speed is still an evolving research problem.

### *2.4. Using a Drone for Human Vital Sign Detection*

The first reported works on using a drone for human breathing and heart rate detection were presented by Al-Naji et al. [12,35]. They used a hovering drone to record the skin colour variation of the facial region at a distance of three meters. Their solution consisted of an improved video magnification technique, spatial averaging, signal decomposition and blind source separation to suppress noise artefacts. The proposed method was constrained to short distance and a single pose where the participant was standing in front of the drone. In another study, Al-Naji et al. [11] proposed a new life signs detector based on motion analysis to efficiently detect alive and deceased humans using a drone at a distances of 4–8 m and in a variety of poses possible in a disaster zone. The results showed that the proposed system can efficiently identify living and non-living human subjects. However, the proposed method requires a human form to be detected. In testing, a deceased subject was represented by a mannequin, which had an appropriate human form. The work presented in our study is an extension to Al-Naji et al.'s life signs detector [11].

## 3. Methodology

This study is focused on analyzing aerial videos from a drone hovering at low-altitude. The first part is video stabilizing. It helps to remove the camera movements. The stabilized video is then divided into tiles as each tile represents a video. Next, magnification was applied on each video (created from tiles) to amplify the frequencies that lie within the human breathing rate range and to attenuate frequencies outside this range. Finally, image differencing and intensity variations were calculated for each video to determine if there was any sign of breathing. The frequency map corresponding to each tiled video provides an estimated location of the human subject. The complete process is shown in a block diagram in Figure 1.

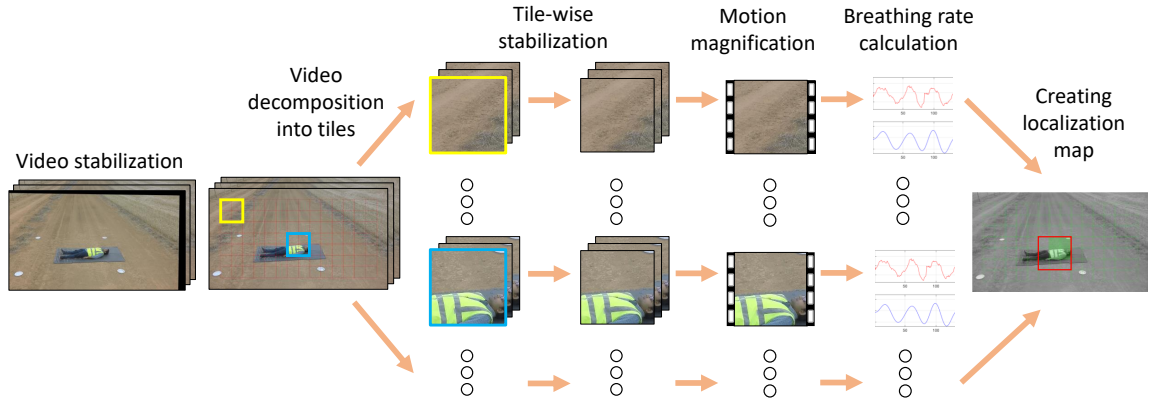


**Figure 1.** Schematic diagram of the proposed approach.

Aerial videos recorded from a drone are inherently dynamic. Despite the use of a stabilised camera platform and the stabilisation of the autopilot, the imagery is constantly in motion. There are also indication of very high frequency movement caused by motors and propellers.

There are several mechanisms in a drone to reduce the motor vibrations experienced by the camera. In our experimental platform, the camera was mounted on a three-axis gimbal. The gimbal attempts to maintain a constant camera attitude and stabilization automatically and achieves stabilized videos within 0.1 degree of pointing accuracy. The vibration of the motors transferred to the camera was minimized by the soft mounts of the gimbal. We used a GoPro Hero 4 black action camera with replacement lens (5.4 mm, 10 MP, infrared IR cut filter) to reduce the fish-eye effect and narrow the field of view. This is a popular camera for aerial and sports activity recording. Despite having these motion reduction mechanisms in place, motion artefacts and drift were present in the low-altitude videos. We address this issue by software stabilizing the videos at two levels, using the similarity transformation.

### *3.1. Video Stabilization*

Our video stabilization technique was inspired by the MathWork's publicly available video stabilization work [13]. We built our algorithm by extending Matlab's code to suit a two level video stabilization for the raw video and video tiles and we adopted similarity transformation instead of the original affine transformation. The stabilization technique is presented as an algorithm in Algorithm 1. An upper limit of 10,000 is set to the maximum number of points to limit computational load. *pThresh* is a scalar representing the minimum intensity difference between a corner and its surrounding region. As *pThresh* increases the number of detected key points decreases. We also set a lower limit for the number of key points. If there are not enough key points for the transformation *pThresh* is increased until a minimum of 10 points are detected. The transformation matrix can be calculated with a minimum of 3 matching points. However, depending of the texture some images do not meet this requirement. In such cases, we use the transformation matrix calculated for the

previous frame. The MSAC algorithm [36] which is a variant of the RANSAC algorithm is used for the feature matching.

Features from the accelerated segment test (FAST) [37] method was used to detect key points in images. The FAST algorithm scans the circular area around the potential key point and looks for a threshold intensity difference between the center point and the surrounding pixels. The selected threshold $T$ was 0.1. $T$ is a scalar defined as a fraction of the maximum intensity value of the image.

Key points are calculated for each adjacent image frame and the new frame is similarity transformed to compensate for the distortion caused by camera movement [38]. We do not perform an affine transform as it warps the new image suppressing the desired human body movements. In similarity transform, the image is possibly uniformly scaled, translated and rotated while preserving the shape details in the image. Figure 2 illustrates the steps involved in the video stabilization step. This figure shows the stabilization of the tenth image frame with respect to the first image frame. In our experiments, the images were stabilized with respect to the previous image. Randomly selected images from a stabilized video are shown in Figure 3. The unfilled areas were created by the transformation.

**Algorithm 1.** Video Stabilization.

**Input**: image $\mathbf{I}\ (n)$, $\mathbf{I}\ (n+1)$;
**Initialization**:
  number of points in $\mathbf{I}\ (n)$, $pointsA$;
  minimum intensity difference, $pThresh$;
**Transformation**:
  **while** $pointsA > 10000$
    detect FAST features of $\mathbf{I}\ (n)$ and $\mathbf{I}\ (n+1)$;
    $pThresh = pThresh + 0.1$:
  **end**
    **for** each $(n+1)$ image
      **while** $pointsA < 10$
        detect FAST features of $\mathbf{I}\ (n)$ and $\mathbf{I}\ (n+1)$;
        extract features in $\mathbf{I}\ (n)$ and $\mathbf{I}\ (n+1)$;
        match features between $\mathbf{I}(n)$ and $\mathbf{I}\ (n+1)$;
        derive similarity transform matrix between $\mathbf{I}\ (n)$ and $\mathbf{I}\ (n+1)$;
        transform $\mathbf{I}(n+1)$;
        $pThresh = pThresh - 0.01$:
      **end**
      reset $pThresh$ to initialized value
    **end**

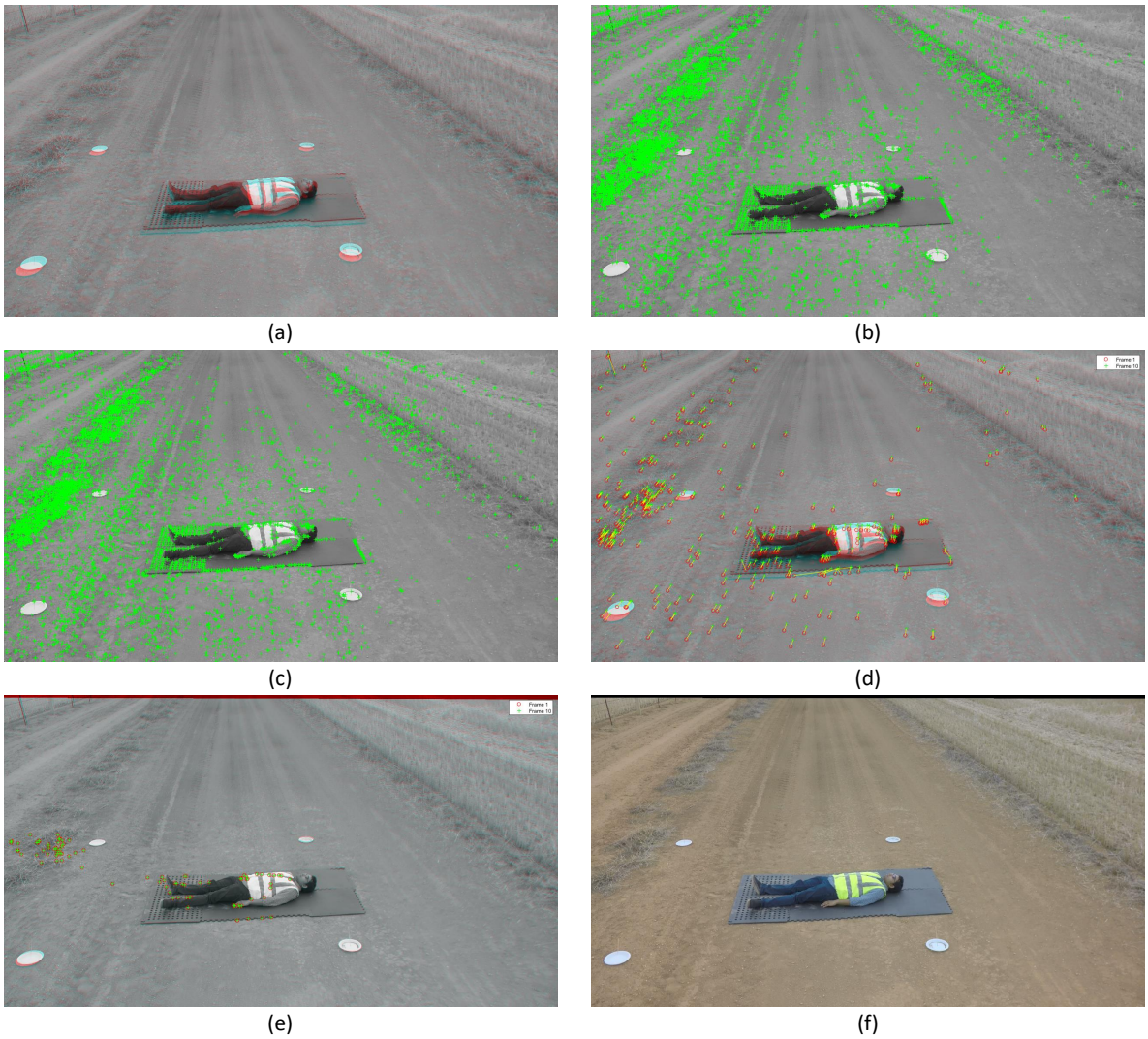


**Figure 2.** Video stabilization steps of the raw video. Stabilization between the image frame 1 and 10 are illustrated here for the demonstration purpose as they have a noticeable offset. However, in the experiments, adjacent images are stabilized. (**a**) Both images are overlaid to show the offset in the frames in red and cyan colors; (**b**,**c**) are the first and tenth frames of the video respectively with their detected key points; (**d**) The matching points between the two images are shown in red and green markers; (**e**) Outlier points are removed and only inlier points are used for the similarity transform; (**f**) The stabilized image.

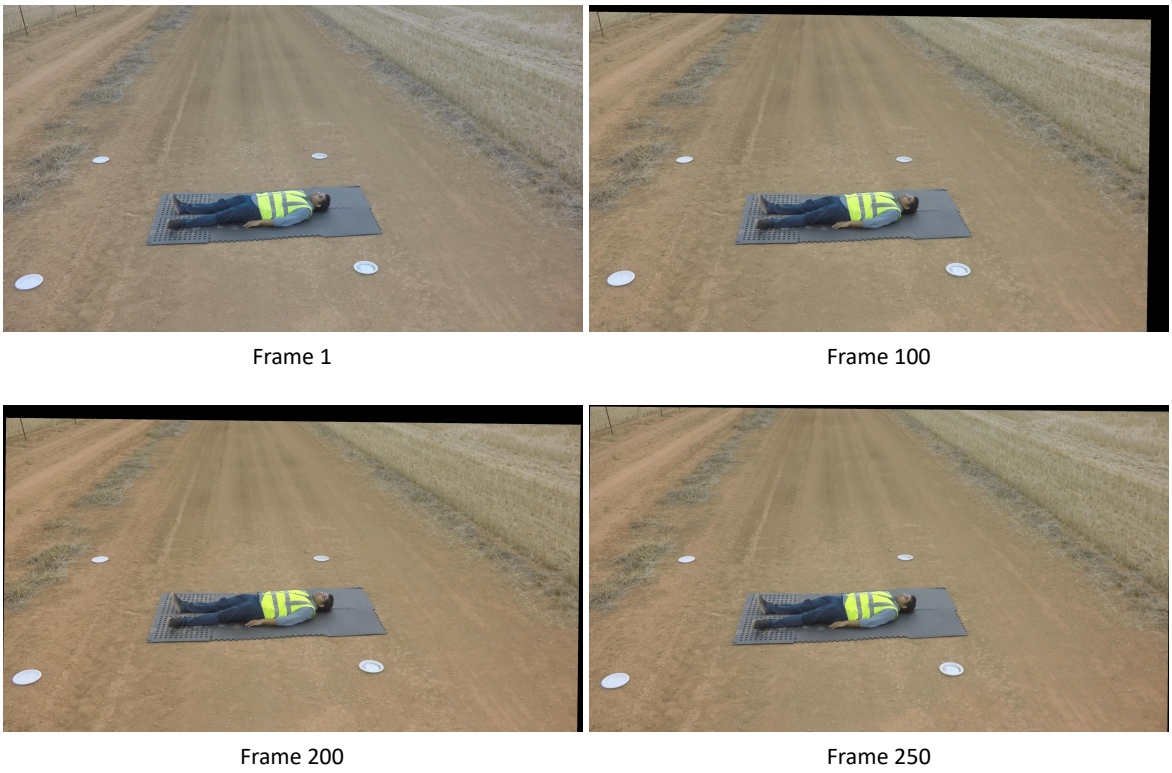


**Figure 3.** Random frames from the stabilized video A1 (video names are given in Table 1). Black colored areas represent the missing areas after the image transformation.

### *3.2. Creation of Image Tiles*

In order to detect the human subject in the image by detecting the breathing rate, we scanned the entire image using a sliding window approach. In this step, the stabilized images were divided into tiles as shown in Figure 4. The resolution of the raw images was 3840 × 2160 pixels. A 240 pixel margin was maintained to allow for unfilled areas generated by the video stabilization step. The size of a tile was 480 × 480 pixels. There were total of 78 tiles per image (6 rows and 13 columns).

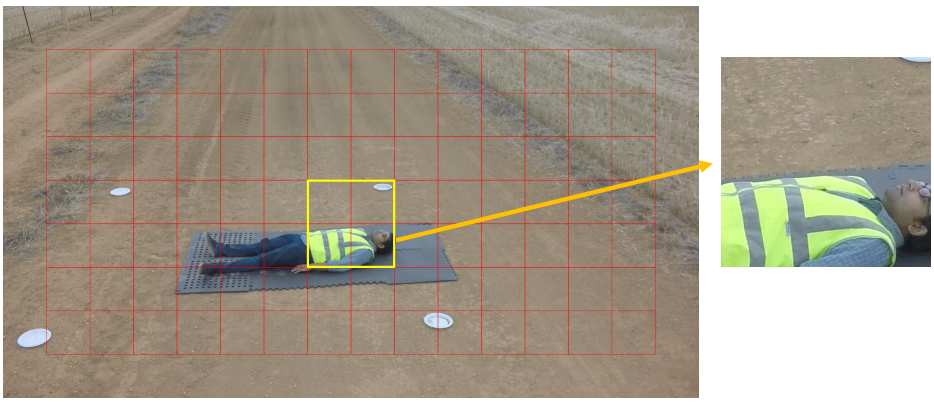

**Figure 4.** The stabilized video is divided in to tiles as shown in the figure. Each yellow coloured box indicates a tile. The red grid shows the locations where the tiles were selected using a sliding window approach. Each tile has 50% overlap with neighbouring tiles in its row and column. The extracted tile from the yellow colour box are shown on the right.

### *3.3. Tile-Wise Video Stabilization*

In the first step, we stabilized the raw video using the key points of adjacent images. However, there can be small camera movements still present in the tile images. To further stabilize the tile images, we ran another video stabilization step on each tile image. The minimum number of points allowed in this step was five. We heuristically defined that a minimum of five points were enough as the image is relatively small. We experimented with both affine and similarity transforms, and selected the similarity transform as it gave less error. A minimum of three matched pairs of points [39] are enough for the similarity transformation. However, higher numbers of matched pairs of points can result in higher accuracy in the estimated transformation. To generate a large number of points we set the initial *pThresh* to be 0.05. If there were not enough points detected we successively decreased *pThresh* down to 0.01 until the minimum number of points are achieved. However, there were some images without any detectable points. They were the images with plain texture. No stabilization was applied to such images.

### *3.4. Creating Videos from Image Tiles*

Video stabilization creates small unfilled areas in the edges of the tile image. Therefore, each image was cropped with a 48 pixel margin (0.1 $\times$ width) to remove the unfilled areas. The cropped images were used to create a new set of videos corresponding to each series of tile images—a total of 78 tile videos.

### *3.5. Motion Magnification*

As the tiles videos were originally extracted from an unstable video, some noise and undesired motion artefacts can still be present. At this point, the desired chest movements can be observed from the tile videos. We do not use any body part detector or manual region of interest (ROI) selection to process the chest area. Instead, we treat the entire tile image as our ROI. This approach is important to detect the breathing rate of an occluded subject or a subject against challenging texture (e.g., covered with dust or wearing camouflage). When processing the whole image we have to deal with a number of object movements in the image caused by different sources. A few examples are swinging plants and grass, flying objects like leaves, random twitches in the subject's clothes caused by wind and camera movements that are residual in the stabilization step. These undesired movements need to be suppressed and the desired chest movement should be amplified. The standard Fourier transform approach cannot be used here as it works only with global phase changes in images [40]. For this requirement, Wadhma et al.'s phase-based motion magnification method [10] which is capable of handling local phase changes was used. We used the publicly available phase-based magnification code without any modification.

Fourier shift theorem can measure global motions. To measure the local phase motions of an image a complex steerable pyramid [41,42] was used. It is similar to a localized Fourier transform that divides the image into spatial structures. A complex steerable pyramid decomposes the image into different sub-bands corresponding to different image scales and orientations. Every pixel location of each sub-band is represented with a local amplitude $A$ and local phase $e^{i\phi}$. The phase signal over time was further processed with a temporal band-pass filter to extract the motion of interest of the pixel location. These temporally band-passed phases are multiplied by amplification factor $\alpha$. The amplified phase differences are used to modify each coefficient of the steerable pyramid. After the phase-based motion processing, all the sub-bands are combined to reconstruct a motion magnified video. Motion magnification can be mathematically explained as follows:

The motion magnified signal of $f(x+\delta(t))$ can be expressed as $f(x+(1+\alpha)\delta(t))$, where $f$ is the image intensity profile and $\delta(t)$ is the initial motion. Using Fourier series decomposition, $f(x+\delta(t))$ can be written as a sum of complex sinusoids where each band corresponds to a single frequency $\omega$.

$$f(x+\delta(t)) = \sum_{\omega=-\infty}^{\infty} A_\omega e^{i\omega(x+\delta(t))}. \tag{1}$$

The frequency band $\omega$ is the complex sinusoid

$$S_\omega(x,t) = \sum_{\omega=-\infty}^{\infty} A_\omega e^{i\omega(x+\delta(t))}. \tag{2}$$

$S_\omega$ is a sinusoidal signal and its phase $\omega e^{i\omega(x+\delta(t))}$ contains motion information. We can modify this phase information to isolate the motion of interest.

This can be done by applying a temporal filter and a DC balanced filter to the phase $i\omega(x+\delta(t))$. Assuming the temporal filter only removes the DC component $\omega x$, the resulting band-pass phase is

$$B_\omega(x,t) = \omega\delta(t). \tag{3}$$

A complex sinusoid $\hat{S}_\omega(x,t)$ that has motions exactly $1+\alpha$ times the input can be obtained by multiplying $B_\omega(x,t)$ and $\alpha$.

$$\hat{S}_\omega(x,t) = S_\omega(x,t)e^{i\alpha B_\omega}, \tag{4}$$

$$\hat{S}_\omega(x,t) = A_\omega e^{i\omega(x+(1+\alpha)\delta(t))}. \tag{5}$$

The motion magnified sequence $f(x+(1+\alpha)\delta(t))$ is calculated by summing all the sub-bands.

We heuristically selected $\alpha$ to be 5 for videos A1-A6 and 8 for videos A7-A10. The interesting frequency band for breathing detection was the range 0.1 – 0.4 Hz.

### *3.6. Calculating the Breathing Rate*

In motion magnified video, the interesting frequency band is amplified and the other frequencies are attenuated. This video shows clearly identifiable image structure changes in potential human torso areas. This phase manipulated video can be further processed with statistical image processing techniques to identify the tiles with potential breathing rates.

First, the images were converted to the frequency domain and their absolute difference was calculated. Image differencing was done between the images $I(t)$ and $I(t+d)$. An image distance of $d$ was selected to obtain a difference image with sufficient detail. The frame frequency of the test videos was 25.1 Hz. With this frame rate, the intensity difference between adjacent images does not provide noticeable pixel changes.

The luminance of pixel values over the difference image sequence was averaged to obtain a representation of the temporal motion changes as

$$I_{avg}(t) = \frac{\sum_{x,y} I_{x,y,t}}{P}, \tag{6}$$

where $I_{x,y,t}$ is the luminance pixel value at location $(x,y)$ over time $t$ and $P$ is the image area.

We de-noised the temporal signal by applying the wavelet signal denoising method [43,44]. It uses an empirical Bayesian estimator with a Cauchy prior. In the denoising step, the signal decomposition is calculated for $N$ levels using soft thresholding. A moving average filter was used to smooth the signal. The number of peaks in the smoothed signal provides an estimate of breathing rate. The intensity signal and the de-noised signals are shown in Figure 5.

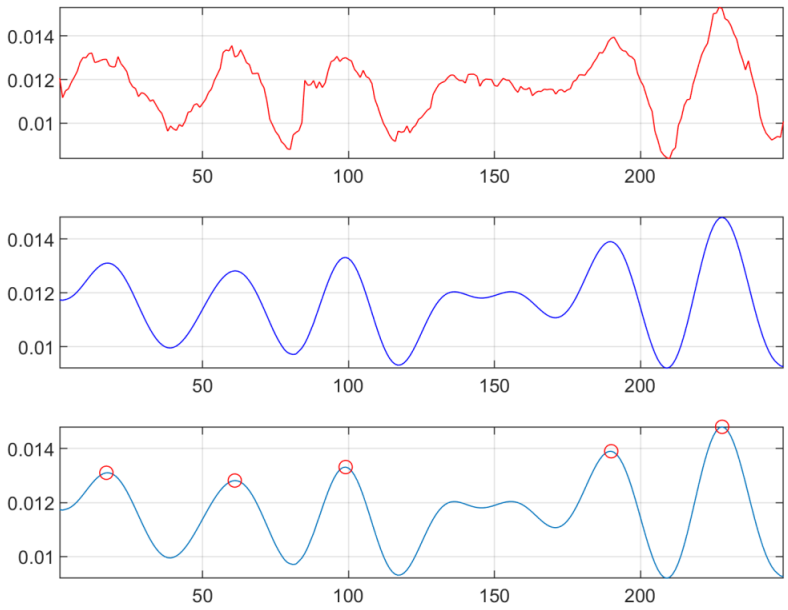


**Figure 5.** The extracted signal from the video A1. It corresponds to the tile shown in Figure 4. The first signal is the temporal intensity changes in the difference image sequence. The band-passed and smoothed signal is shown in the middle and the detected peaks are marked on the third signal. x and y axes represent the number of frames and the intensity difference respectively.

### *3.7. Creating a Localization Map*

Breathing rate estimations were calculated for all tile videos. We use these estimates to create an overlay map with potential human subject locations. Each estimate was mapped to their corresponding image locations. Estimates that fall inside the range of human breathing rate (12–20 breaths per minute) were presented as a colored grid map over the first image frame of the video. Resulting localization maps are provided in the next section (see the second column of Figures 6–8).

### *3.8. Comparison to the State-of-the-Art*

Our technique uses the movement associated with life signs to detect the presence of humans in aerial video. This requires a combination of image stabilisation and signal processing of image changes to detect the specific band of breathing frequencies. We developed a technique to analyse the breathing rate from drone video without regard for the human form, color or intensity. This is a question of analysing small cyclic motions (breathing) in the presence of large mostly random motions (drone and camera movement). The closest state-of-the-art image processing technique available in the literature to address part of this question is the work presented by Elgharib et al. [45]. Here, we qualitatively compare our approach to theirs.

Their work focuses on small motion magnification in the presence of large motions. For a range of videos, their method has shown state-of-the-art results. However, there are limitations in the approach in the context of drone video analysis for human detection.

Elgharib et al.'s method consists of two main modules—video stabilization and motion magnification. In the video stabilization module, the image sequence is stabilized over a reference frame. In the magnification module, manual interaction is necessary. The user should select the region of interest (ROI) and then video magnification is applied only on the selected foreground object. Their proposed method is a general object focused approach and ours is an application specific whole image approach.

## 4. Experimental Results

Aerial videos were recorded using two drones—a 3DR Solo and a DJI Mavic Pro. A GoPro Hero 4 Black action camera with replacement lens (5.4 mm, 10 MP, IR CUT) to reduce the fish-eye effect and narrow the field of view was used with the 3DR Solo drone. The GoPro camera was mounted on a 3-axis camera stabilising gimbal. The pointing accuracy of the gimbal was 0.1 degrees. In both drones, the videos were sampled at 25 frames per second and the resolution was $3840 \times 2160$ pixels. During the data capture, the drone was in hovering state at 4–8 m altitude.

The data collection was conducted under the approval of University of South Australia's Human Research Ethics Committee (protocol no. 0000035185). Six participants were involved and data capture was done on multiple days at different locations. We selected a relatively clutter free background and a complex background for video recording. The participants were asked to lie down comfortably and to breath naturally pretending to sleep. The videos were recorded at different angles. Five videos of fully visible subjects and one video of head and leg covered subject were selected for the clutter free experiments (see Figure 6). These videos were recorded using the 3DR Solo drone. We also collected four videos of a human subject covered with a camouflage net using the Mavic Pro drone (see Figure 7). The camouflage net and the complex background simulated a real-life scenario. The simulated scenarios were (i) a camouflaged person and clearly visible person, (ii) a person covered with a camouflage net, (iii) a camouflaged person and a mannequin and (iv) a camouflaged person and mannequin partly covered with sheet metal.

To compare our method with the stable videos, we also collected videos from a stationary camera. We selected three videos with a single subject on each and one video with a subject lying down next to a mannequin. The mannequin was a 1.96m tall male mannequin with a realistic face. It was fully clothed and wearing a black wig.

The entire solution was implemented in MATLAB. The video stabilization and the motion magnification are the computationally heavy components of the experiment. The video stabilization of a 30 s original video (4k resolution) took approximately 12 min. Video stabilization and motion magnification of a 30 s tile video ($480 \times 480$ resolution) took between 3 and 6 min. The algorithms were implemented on a core i7, 2.6 GHz laptop computer using an unoptimized MATLAB scripts. The scope of this investigative study was limited to offline data processing. We did not design the experiments for onboard processing or real-time video transmission even though they are vital components in practical SAR image analysis.

We analyzed the performance of our estimation using the overlap [46]. The overlap measure is often used in object tracking experiments to determine how well the estimated bounding box matches the ground truth bounding box. The overlap measure in our experiment was computed by taking the ratio between the intersection region and union region of ground truth bounding boxes and estimated bounding boxes. If the regions covered by ground truth bounding boxes and estimated bounding boxes are $A_g$ and $A_e$ respectively, the overlap can be measured from

$$F(A_g, A_e) = \frac{A_g \cap A_e}{A_g \cup A_e}. \tag{7}$$

We also calculated the measurement precision of the estimations. Precision is the the ability of a classification model to identify only the relevant data points, that is, the fraction of detected items that are correct. It is defined as

$$Precision = \frac{\{True\ Positives\}}{\{True\ Positives\} + \{False\ Positives\}}. \tag{8}$$

Both overlap and precision are widely used measures in machine learning object detection problems. In this study, we address an object detection problem but our solution is an analytical

approach rather than a machine learning-based numerical approach. We selected the above performance measures as they represent the object detection accuracy well irrespective of the approach.

Image tiles that covered the chest area of the frame of the sequence were considered to be the ground truth bounding boxes. We assume that the change in location of the chest area in all frames with respect to its location in the first frame was negligible. We report the experimental results in Table 1.

**Table 1.** A summary of the selected videos and the estimation accuracy. We named the aerial videos from A1 to A6 and stationary videos from S1 to S4. Here, MM Fcy refers to motion magnification frequency. "Scenario" column lists additional remarks to the lying down pose. These results are graphically illustrated in Figures 6 and 8.

| Video Name | Scenario | #Frames | MM Fcy (Hz) | Overlap | Precision | Remarks |
|---|---|---|---|---|---|---|
| A1 | face up | 253 | 0.3–0.4 | 0.45 | 1 | successful |
| A2 | face down | 752 | 0.3–0.4 | 0.38 | 0.46 | partially successful |
| A3 | face down | 753 | 0.3–0.4 | 0.36 | 0.67 | partially successful |
| A4 | face up, human & mannequin | 725 | 0.3–0.4 | 0.67 | 1 | successful |
| A5 | face & legs covered | 752 | 0.3–0.4 | 0.67 | 1 | successful |
| A6 | face down | 760 | 0.3–0.4 | 0.36 | 1 | successful |
| A7 | camouflaged | 751 | 0.2–0.3 | 0.47 | 1 | successful |
| A8 | camouflaged | 751 | 0.2–0.3 | 1 | 1 | successful |
| A9 | camouflaged | 751 | 0.2–0.3 | 1 | 1 | successful |
| A10 | camouflaged | 751 | 0.2–0.3 | 1 | 1 | successful |
| S1 | face up, human | 759 | 0.3–0.4 | 1 | 1 | successful |
| S2 | face up | 582 | 0.1–0.3 | 1 | 1 | successful |
| S3 | face up | 534 | 0.1–0.3 | 1 | 1 | successful |
| S4 | face up | 766 | 0.1–0.3 | 1 | 1 | successful |

The average overlap values of aerial videos and ground recorded videos are 0.42 and 1, respectively. The main reason for not achieving a higher overlap was motion artefacts. Our video stabilization step compensates for most of the camera motions but it is not perfect. As the videos were recorded at low altitude (4–8 m) it is possible that wash from the rotor contributed to swinging motion of the grass in the background particularly for the A2 video.

Precision is a better representation of performance than overlap for our application. We do not need precise bounding box locations as our intention is to locate the person in the image. Therefore, we used precision as the performance measure in our experiment. The average precision of aerial videos and ground videos was 0.913 and 1, respectively. Videos of a camouflaged person (A7–A10) show higher detection accuracy compared to the clutter-free videos (A1–A6).

The overall detection correlates well with the ground truth. All ground record videos showed perfect detection accuracies. Since there were no motion artefacts the image structure changes from chest movement was clearly detectable. Although, the widely accepted breathing rate range is 12–20 breaths/min or 0.2–0.33 Hz, we noticed that some participants had breathing rates outside this range in the recorded videos. Therefore, for motion magnification, we used three frequency ranges. A discussion about the limitations in this approach inducing imperfect stabilization is provided in Section 5.

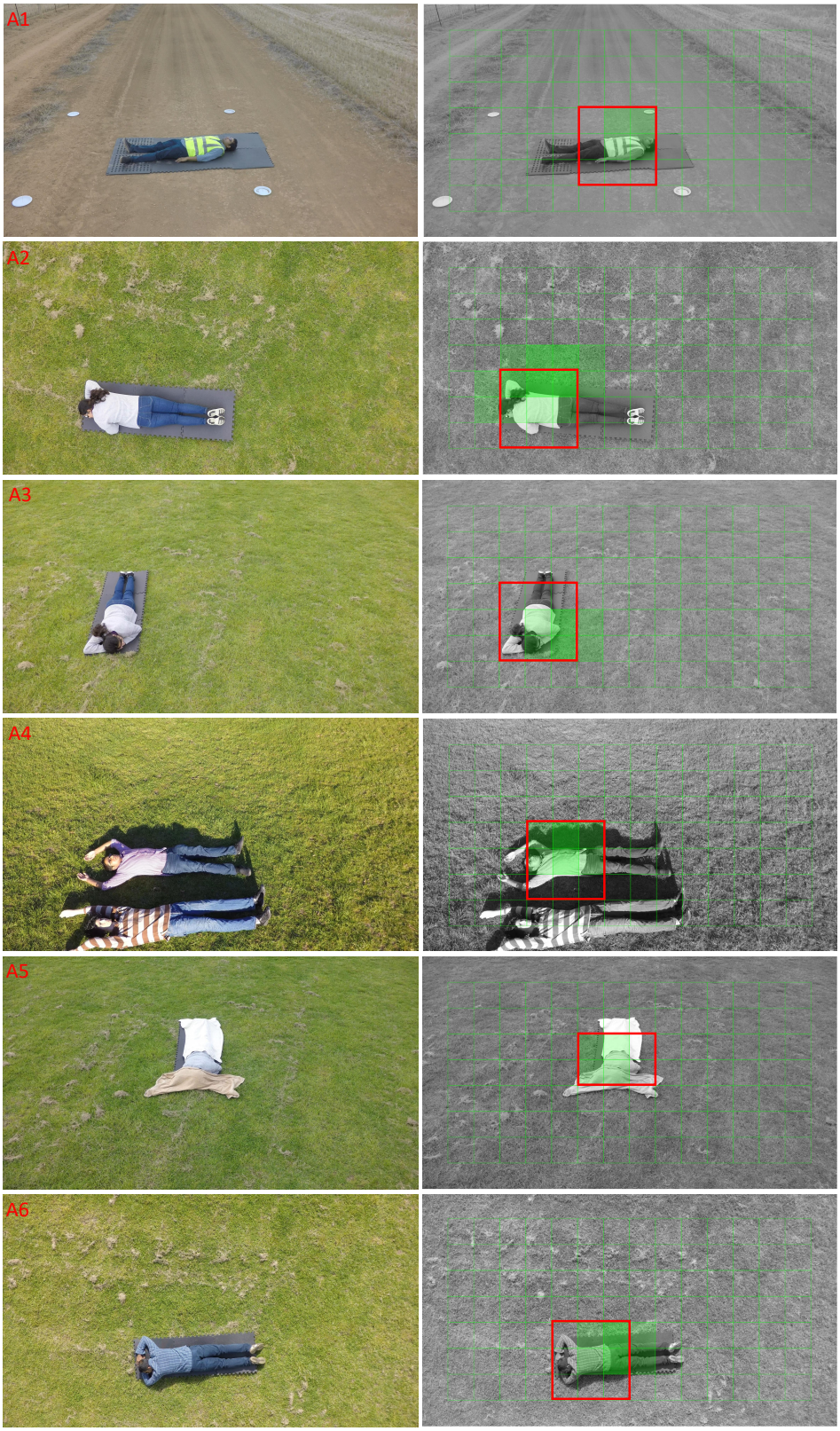


**Figure 6.** The first column of the images are the first frames of each aerial video (A1–A6). Aerial videos A7–A10 are shown in Figure 7. The second column of images show their localization maps drawn over the first frame. Ground truth bounding boxes are shown in red. Detected bounding boxes are colored green.

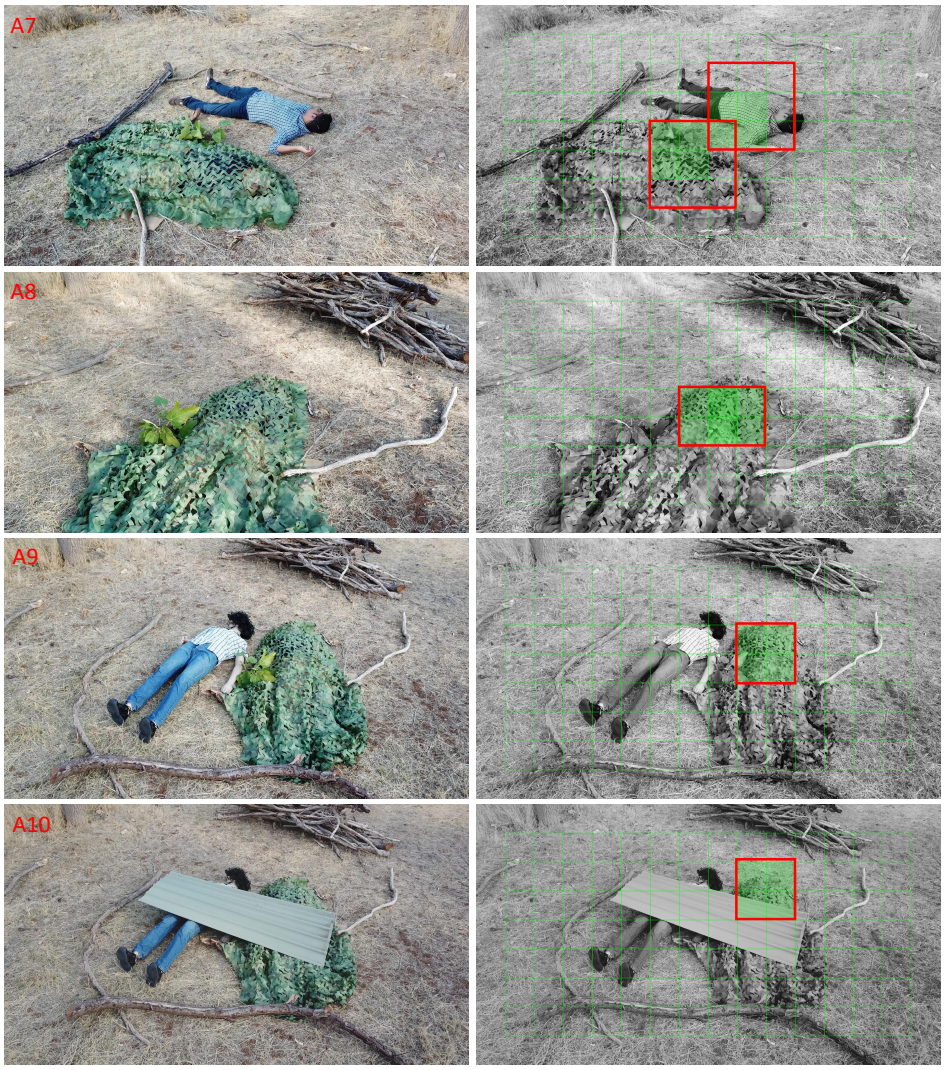


**Figure 7.** Aerial videos A7–A10. In these videos, the human subject was covered with a camouflage net. The simulated scenarios were A7: a camouflaged person and clearly visible person, A8: a person covered with a camouflaged net, A9: a camouflaged person and a mannequin and A10: a camouflaged person and mannequin partly covered with a sheet metal. A1–A6 are shown in Figure 6.

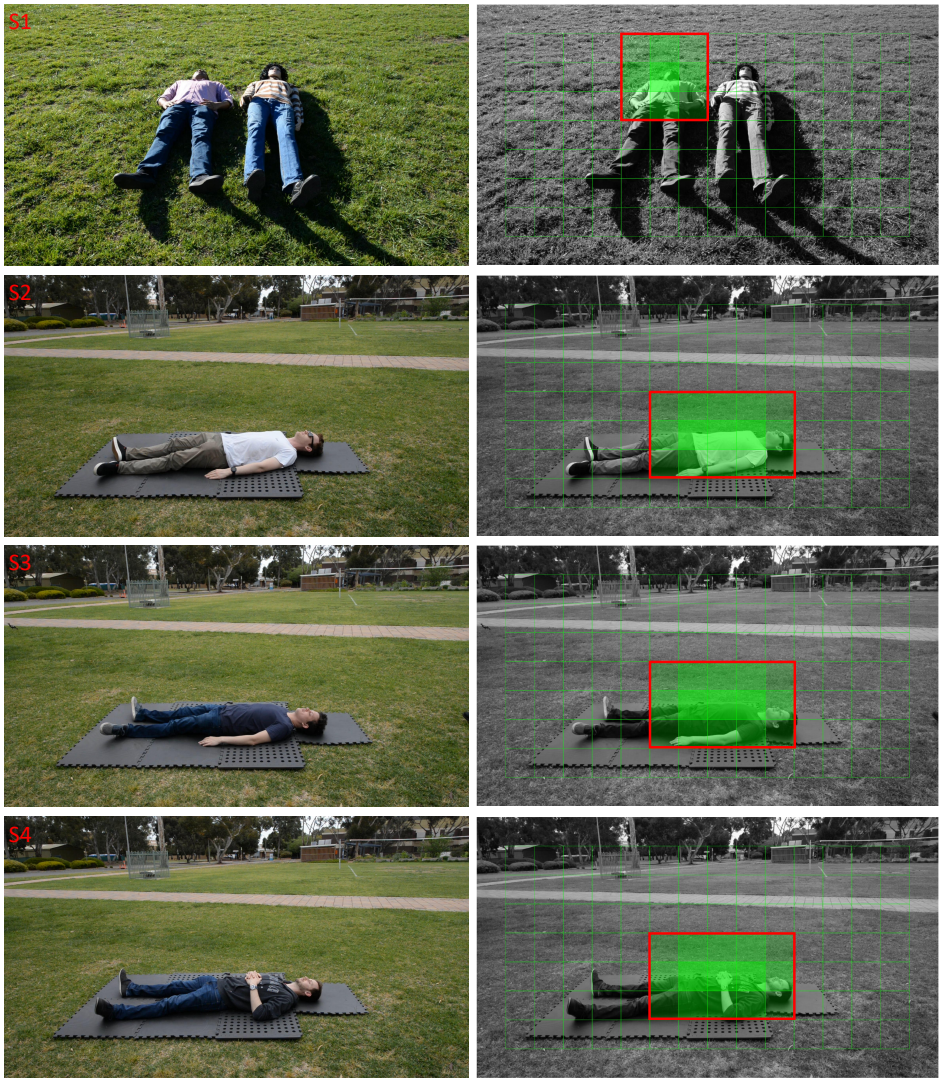


**Figure 8.** The first column of the images are the first frames of each ground video. The second column of images show their localization maps drawn over the first frame.

## 5. Discussion

This approach was proposed to detect live human subjects from aerial video. As the existing techniques focus on the human shape or body heat those techniques might not work in challenging conditions like detecting occluded live people in broad daylight. We tested the proposed method with aerial as well as ground videos recorded from a stationary camera. There were some false detections in the aerial videos caused by motion artefacts but the ground based videos gave perfect detection results.

Our study consists of both theoretical and experimental components and it has not been applied as an emergency tool before. We are confident that this work has the potential to further develop to detect camouflaged or completely covered live humans from aerial videos and to be used as an emergency tool by enhancing the hardware and software components. The work adds impetus to the problem of image stabilization.

The main limitation of this approach is computational complexity. As the stabilized videos is decomposed into tiles, the process needs to be repeated for tile videos. However, by combining this approach with a human detector the computational cost can be reduced by approximately 90%. When the human location is known only the tiles around the chest area need to be processed to detect a potential heart rate. However, a human detector does not work for unidentifiable human shapes whereas our method produces promising results in these videos. The next limitation is imperfect video

stabilization. The stabilization function can be further improved at a higher computational cost. In this study, we tried to maintain a compromise between computational complexity and the accuracy for two reasons: (i) this is the first study of this kind, to our knowledge; and (ii) we wanted to validate our technique.

Standard video stabilization methods use projective transformation [47], affine transformation [32] or similarity transformation [48], or combinations of them [49,50] as their 2D motion model. However, our initial tests with projective transformation did not give accurate results. We also conducted more tests with affine and similarity transforms. For two stage stabilization, we used the combinations of affine-affine, affine-similarity, similarity-affine and similarity-similarity. Our test results showed that similarity-similarity combination gave better results compared to other combinations. Therefore, in this study we report the similarity transformation as our 2D motion model. Figure A1a shows the raw input mean and corrected sequence mean of video A1. The corrected sequence mean is calculated after the first level of stabilization. In the second level of stabilization, individual tiles are further stabilized using the similarity transform. The raw input mean and corrected sequence mean of randomly selected tile videos are shown in Figure A1b. Our observation was that the key point distribution pattern of the video was the reason for getting better results for similarity transform compared to projective and affine transformations. In our test videos, the matched pairs of points are mostly concentrated around the object of interest. These key points dominate the projective and affine transformations and change the shape of the object of interest during the transformation. However, a uniform key point distribution or collecting of key points from evenly distributed areas will facilitate using projective or affine transforms with less errors. As we have designed this preliminary study to prove the concept, we have limited our test to a controlled setting and the described video stabilization technique. This observation warrants the need for real-world test videos and further improvements to our stabilization step.

Video stabilization cannot be completely achieved by image processing techniques. There are certain improvements to be done during the data capture. For data capture, we used a 4K resolution camera. The consumer grade camera used in this study produces high levels of noise. In order to enhance the video stabilization step more image details could be captured. A higher resolution camera (e.g., 8K resolution) should facilitate finer stabilization. The drone platform (3DR Solo) we used for data collection was relatively light with 1.8 kg of total weight and relatively dated in design. Lighter drones are susceptible to wind gusts when they are in a hovering state at a GPS-locked location. Movement caused by wind changes the geometry captured by the camera. To avoid such undesired movement, a more stable drone could be used.

There are hardware limitations with drone cameras. Conventional cameras have an angular/spatial coverage limited by their optics. In this application an alternate arrangement might be desirable. A possible but expensive solution is to use a custom-made camera system that captures a much larger angle on multiple focal planes simultaneously.

We noted that when the drone hovers down to 4m altitude, grass and small plants start to sway due to rotor wash. These movements resulted in some false positive detections. We maintained our recording platform altitude at 4–8 m. When the drone hovers to a lower altitude it captures more image details but suffers more wash. Conversely, higher altitudes have minimal or no wash affects on the subject but distance limits the image details. Therefore, a sufficiently high altitude and high image resolution are recommended. A higher resolution camera combined with a stable drone should facilitate much more stable videos for similar experiments. Another option might also be simply to stare at the same location for longer to firm up the statistics of detection. A question raised by the image stabilisation problems is that maybe the drone gimbal is not really necessary and might even be causing motion artefacts of its own.

Our experiments showed that videos recorded perpendicular to the ground gave inaccurate result as the chest movements were towards and away from the camera. These movements had an

insignificant structure change in the image. Videos captured at an angle showed better results as their chest movements were vertical in the image plane and relatively easy to detect.

The widely accepted breathing rate for an adult is 12–20 breaths per minute. However, this can change depending on the physique of the person and the situation. We have not explored the possibility of longer term observations to further isolate the breathing signal. In principle this is a steady cyclic motion that is not correlated with wind, rotor wash or motion of the drone. With time, the statistics of the signal should emerge from the other random signals, perhaps with even less image stabilisation, as long as the mean location of the drone and angle of the camera stays the same.

The videos of a camouflaged person were recorded in a complex background using the more stable Mavic Pro drone. Therefore, the key point-based video stabilization gave more accurate results for complex backgrounds compared to the clutter-free backgrounds. In the motion amplified tile videos, the chest movement of the camouflaged person was clearly visible. This experiment proved the applicability of our method in complex and challenging scenarios.

Our immediate plan is to extend this solution to more realistic scenarios. This solution was tested in a controlled environment and it is a computationally expensive process. The video stabilization and motion magnification components do not run in real-time. This is a pilot study and it is a first step towards detecting living people by observing movement associated with breathing. We have collected videos when the humans are lying down on a flat ground. However, in realistic search and rescue scenarios, more complex backgrounds and partially and fully covered humans should be expected. Use of a thermal camera in parallel to the RGB camera will help to address such challenging scenarios. The thermal profile of the body can be used to locate the humans and then analyze their breathing. Our method in its current form will not work in windy conditions as the wind could easily create undesired motions in surroundings and clothing and prevent the breathing movement being identified by the camera. In this study, we have proved that the proposed method is working under controlled conditions and it has the potential to extend for real-life scenarios with more enhancements. Moreover, for possible improvements, we can leverage the state-of-the-art technologies of the following areas: (i) identifying living people using a drone when their body is clearly visible; (ii) identifying living people using a drone when their body is partially or fully covered; (iii) video stabilization in drone videos; and (iv) contactless breathing analysis algorithms for real-time operations.

## 6. Conclusions

In this study, we proposed a method to detect live humans from aerial videos. Our approach does not need to detect the human body shape in whole or in part to detect live subjects. Instead it detects image motion near breathing rates that are detected from aerial video. There are four main components to this approach: (i) video stabilization using the key points of adjacent images; (ii) video decomposition into tiles to separately analyze for possible breathing rate detection; (iii) motion magnification to amplify potential chest movements and to attenuate unwanted frequency bands; and (iv) image differencing band-pass filtering to calculate the breathing rate in each tile video. The video tiles with an estimated breathing rate are mapped to their original locations in the image frame and are presented as a localization map.

We experimented with aerial as well as ground recorded videos. The ground recorded videos showed perfect results using our method. We showed that with good video stabilization it is possible to detect live human subjects from their breathing. The localization map helps to detect live people against and among challenging textures where other proposed techniques fail. Some potential examples are detecting camouflaged people or people covered with dust or ash. This is an investigative, pilot work on the feasibility of a system like this, with further evaluation on real-world data being necessary to establish how to achieve a practical SAR system. Our ongoing work includes extending this method to more challenging and realistic textures with possible reduction of motion artefacts and computational complexity. Possible future directions of this work are object-centered video

stabilization, algorithm optimization to achieve a reduced computational complexity and GPU implementation for faster processing.

**Author Contributions:** Conceptualization, J.C. and A.G.P.; methodology, A.G.P., F.-T.-Z.K.; investigation, A.G.P.; resources, J.C.; data curation, A.G.P.; writing—original draft preparation, A.G.P., F.-T.-Z.K.; writing—review and editing, A.A.-N, J.C.; supervision, J.C.; Software, A.G.P. and A.A.-N. All authors have read and agreed to the published version of the manuscript.

**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. We acknowledge Tran Nguyen for his technical support in data collection. We also thank MIT Computer Science and Artificial Intelligence Lab for making their motion magnification codes publicly available.

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

## Appendix A. Mean Images of Video Stabilized Image Sequence A1


**Figure A1.** Some selected mean images after the video stabilization of video A1. (**a**) The raw input mean (left) and corrected sequence mean (right) of video A1. The corrected sequence mean is calculated after the first level of stabilization. (**b**) Mean images after the second level of stabilization. Each pair of images represents the raw input mean (left) and corrected sequence mean (right) of a randomly selected tile video of A1.

## References


1. Waharte, S.; Trigoni, N. Supporting search and rescue operations with UAVs. In Proceedings of the 2010 International Conference on Emerging Security Technologies, Canterbury, UK, 6–7 September 2010; pp. 142–147.
2. Murphy, R.R.; Tadokoro, S.; Kleiner, A. Disaster robotics. In *Springer Handbook of Robotics*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 1577–1604.
3. Liu, Y.; Nejat, G. Multirobot cooperative learning for semiautonomous control in urban search and rescue applications. *J. Field Robot.* **2016**, *33*, 512–536. [CrossRef]
4. Liu, Y.; Ficocelli, M.; Nejat, G. A supervisory control method for multi-robot task allocation in urban search and rescue. In Proceedings of the 2015 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR), West Lafayette, IN, USA, 18–20 October 2015; pp. 1–6.
5. Doroodgar, B.; Liu, Y.; Nejat, G. A learning-based semi-autonomous controller for robotic exploration of unknown disaster scenes while searching for victims. *IEEE Trans. Cybern.* **2014**, *44*, 2719–2732. [CrossRef] [PubMed]
6. Beloev, I.H. A review on current and emerging application possibilities for unmanned aerial vehicles. *Acta Technol. Agric.* **2016**, *19*, 70–76. [CrossRef]
7. Mayer, S.; Lischke, L.; Woźniak, P.W. Drones for Search and Rescue. In Proceedings of the First International Workshop on Human-Drone Interaction, Glasgow, UK, 4–5 May 2019.
8. Chen, X.; Cheng, J.; Song, R.; Liu, Y.; Ward, R.; Wang, Z.J. Video-Based Heart Rate Measurement: Recent Advances and Future Prospects. *IEEE Trans. Instrum. Meas.* **2018**, *68*, 3600–3615. [CrossRef]
9. Khanam, F.T.Z.; Al-Naji, A.; Chahl, J. Remote Monitoring of Vital Signs in Diverse Non-Clinical and Clinical Scenarios Using Computer Vision Systems: A Review. *Appl. Sci.* **2019**, *9*, 4474. [CrossRef]
10. Wadhwa, N.; Rubinstein, M.; Durand, F.; Freeman, W.T. Phase-based Video Motion Processing. *ACM Trans. Graph.* **2013**, *32*, 80:1–80:10. [CrossRef]
11. Al-Naji, A.; Perera, A.G.; Mohammed, S.L.; Chahl, J. Life Signs Detector Using a Drone in Disaster Zones. *Remote Sens.* **2019**, *11*, 2441. [CrossRef]
12. Al-Naji, A.; Perera, A.G.; Chahl, J. Remote monitoring of cardiorespiratory signals from a hovering unmanned aerial vehicle. *Biomed. Eng. Online* **2017**, *16*, 101. [CrossRef]
13. Video Stabilization Using Point Feature Matching. Available online: https://www.mathworks.com/help/vision/examples/video-stabilization-using-point-feature-matching.html. (accessed on 1 December 2019).
14. Grogan, S.; Gamache, M.; Pellerin, R. The Use of Unmanned Aerial Vehicles and Drones in Search and Rescue Operations—A Survey. In Proceedings of the PROLOG 2018, Hull, UK, 28–29 June 2018.
15. Andriluka, M.; Schnitzspan, P.; Meyer, J.; Kohlbrecher, S.; Petersen, K.; von Stryk, O.; Roth, S.; Schiele, B. Vision based victim detection from unmanned aerial vehicles. In Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Taipei, Taiwan, 18–22 October 2010; pp. 1740–1747. [CrossRef]
16. Camara, D. Cavalry to the rescue: Drones fleet to help rescuers operations over disasters scenarios. In Proceedings of the 2014 IEEE Conference on Antenna Measurements and Applications (CAMA), Antibes Juan-les-Pins, France, 16–19 November 2014; pp. 1–4.
17. Sulistijono, I.A.; Risnumawan, A. From concrete to abstract: Multilayer neural networks for disaster victims detection. In Proceedings of the 2016 International Electronics Symposium (IES), Bali, Indonesia, 29–30 September 2016; pp. 93–98.
18. Lygouras, E.; Santavas, N.; Taitzoglou, A.; Tarchanidis, K.; Mitropoulos, A.; Gasteratos, A. Unsupervised Human Detection with an Embedded Vision System on a Fully Autonomous UAV for Search and Rescue Operations. *Sensors* **2019**, *19*, 3542. [CrossRef]
19. Bejiga, M.B.; Zeggada, A.; Nouffidj, A.; Melgani, F. A Convolutional Neural Network Approach for Assisting Avalanche Search and Rescue Operations with UAV Imagery. *Remote Sens.* **2017**, *9*, 100. [CrossRef]
20. Al-Kaff, A.; Gómez-Silva, M.J.; Moreno, F.M.; de la Escalera, A.; Armingol, J.M. An appearance-based tracking algorithm for aerial search and rescue purposes. *Sensors* **2019**, *19*, 652. [CrossRef] [PubMed]
21. Yamazaki, Y.; Tamaki, M.; Premachandra, C.; Perera, C.J.; Sumathipala, S.; Sudantha, B.H. Victim Detection Using UAV with On-board Voice Recognition System. In Proceedings of the 2019 Third IEEE International Conference on Robotic Computing (IRC), Naples, Italy, 25–27 February 2019; pp. 555–559. [CrossRef]

22. Portmann, J.; Lynen, S.; Chli, M.; Siegwart, R. People detection and tracking from aerial thermal views. In Proceedings of the 2014 IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China, 31 May–7 June 2014; pp. 1794–1800.
23. Kang, J.; Gajera, K.; Cohen, I.; Medioni, G. Detection and tracking of moving objects from overlapping EO and IR sensors. In Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop, Washington, DC, USA, 27 June–2 July 2004; pp. 123–123.
24. Doherty, P.; Rudol, P. A UAV Search and Rescue Scenario with Human Body Detection and Geolocalization. In *AI 2007: Advances in Artificial Intelligence*; Orgun, M.A., Thornton, J., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; pp. 1–13.
25. Rivera, A.; Villalobos, A.; Monje, J.; Mariñas, J.; Oppus, C. Post-disaster rescue facility: Human detection and geolocation using aerial drones. In Proceedings of the 2016 IEEE Region 10 Conference (TENCON), IEEE, Singapore, 22–25 November 2016; pp. 384–386.
26. Blondel, P.; Potelle, A.; Pégard, C.; Lozano, R. Fast and viewpoint robust human detection for SAR operations. In Proceedings of the 2014 IEEE International Symposium on Safety, Security, and Rescue Robotics (2014), Hokkaido, Japan, 27–30 October 2014; pp. 1–6.
27. Wang, J.; Feng, Z.; Chen, Z.; George, S.A.; Bala, M.; Pillai, P.; Yang, S.; Satyanarayanan, M. Edge-Based Live Video Analytics for Drones. *IEEE Internet Comput.* **2019**, *23*, 27–34. [CrossRef]
28. Zhang, C.; Zhang, W. Spectrum Sharing of Drone Networks. *IEEE J. Select. Areas Commun.* **2017**, *35*, 136–144.
29. Wang, J.; Feng, Z.; Chen, Z.; George, S.; Bala, M.; Pillai, P.; Yang, S.; Satyanarayanan, M. Bandwidth-Efficient Live Video Analytics for Drones Via Edge Computing. In Proceedings of the 2018 IEEE/ACM Symposium on Edge Computing (SEC), Seattle, WA, USA, 25–27 October 2018; pp. 159–173. [CrossRef]
30. Liu, F.; Gleicher, M.; Wang, J.; Jin, H.; Agarwala, A. Subspace Video Stabilization. *ACM Trans. Graph.* **2011**, *30*, 1–10. [CrossRef]
31. Liu, F.; Gleicher, M.; Jin, H.; Agarwala, A. Content-Preserving Warps for 3D Video Stabilization. *ACM Trans. Graph.* **2009**, *28*, 1–9. [CrossRef]
32. Matsushita, Y.; Ofek, E.; Tang, X.; Member, S.; yeung Shum, H. Full-frame video stabilization with motion inpainting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 1150–1163. [CrossRef]
33. Walha, A.; Wali, A.; Alimi, A.M. Video stabilization with moving object detecting and tracking for aerial video surveillance. *Multimed. Tools Appl.* **2015**, *74*, 6745–6767. [CrossRef]
34. Wang, Y.; Hou, Z.; Leman, K.; Chang, R. Real-Time Video Stabilization for Unmanned Aerial Vehicles. In Proceedings of the MVA 2011, Nara, Japan, 13–15 June 2011.
35. Al-Naji, A.; Perera, A.G.; Chahl, J. Remote measurement of cardiopulmonary signal using an unmanned aerial vehicle. *IOP Conf. Ser. Mater. Sci. Eng.* **2018**, *405*, 012001. [CrossRef]
36. Torr, P.; Zisserman, A. MLESAC: A New Robust Estimator with Application to Estimating Image Geometry. *Comput. Vis. Image Underst.* **2000**, *78*, 138 – 156. [CrossRef]
37. Rosten, E.; Drummond, T. Fusing points and lines for high performance tracking. In Proceedings of the ICCV 2005, Beijing, China, 17–20 October 2005; Volume 2, pp. 1508–1515.
38. Wang, Y.; Chang, R.; Chua, T.W.; Leman, K.; Pham, N.T. Video stabilization based on high degree B-spline smoothing. In Proceedings of the 21st International Conference on Pattern Recognition (ICPR 2012), Tsukuba, Japan, 11–15 November 2012; pp. 3152–3155.
39. Mathworks. Geometric Transformation Types for Control Point Registration. 2019. Available online: https://au.mathworks.com/help/images/geometric-transformation-types-for-control-point-registration.html (accessed on 1 December 2019).
40. Wang, Z.; Simoncelli, E.P. Local phase coherence and the perception of blur. In *Advances in Neural Information Processing Systems*; 2004; MIT Press, Cambridge, MA, pp. 1435–1442.
41. Simoncelli, E.P.; Freeman, W.T.; Adelson, E.H.; Heeger, D.J. Shiftable multiscale transforms. *IEEE Trans. Inf. Theory* **1992**, *38*, 587–607. [CrossRef]
42. Portilla, J.; Simoncelli, E.P. A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. *Int. J. Comput. Vis.* **2000**, *40*, 49–70.:1026553619983. [CrossRef]
43. Donoho, D.L. De-noising by soft-thresholding. *IEEE Trans. Inf. Theory* **1995**, *41*, 613–627. [CrossRef]
44. Donoho, D.L.; Johnstone, I.M. Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **1994**, *81*, 425–455, doi:10.1093/biomet/81.3.425. [CrossRef]

45. Elgharib, M.; Hefeeda, M.; Durand, F.; Freeman, W.T. Video Magnification in Presence of Large Motions. In Proceedings of the The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
46. Kristan, M. et al.. The sixth Visual Object Tracking VOT2018 challenge results. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
47. Buehler, C.; Bosse, M.; McMillan, L. Non-metric image-based rendering for video stabilization. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, Kauai, HI, USA, 8–14 December 2001; Volume 2, pp. 609–614. [CrossRef]
48. Veldandi, M.; Ukil, S.; Rao, K.G. Video stabilization by estimation of similarity transformation from integral projections. In Proceedings of the 2013 IEEE International Conference on Image Processing, Melbourne, Australia, 15–18 September 2013; pp. 785–789. [CrossRef]
49. Lee, K.Y.; Chuang, Y.Y.; Chen, B.Y.; Ouhyoung, M. Video stabilization using robust feature trajectories. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 1397–1404. [CrossRef]
50. Grundmann, M.; Kwatra, V.; Essa, I. Auto-directed video stabilization with robust L1 optimal camera paths. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; pp. 225–232. [CrossRef]