ORIGINAL RESEARCH **Open access**

Enhancing Person Detection in UAV Image Sequences via Displacement Vector Modeling

Maria Silva[1*], Joao Pereira[1], Bruno Costa[2], Ricardo Alves[3]

Abstract

Latest advancements in detecting individuals using deep learning techniques on aerial photographs taken by Unmanned Aerial Vehicles (UAVs) have demonstrated the effectiveness of this approach in contexts such as Search and Rescue (SAR) missions. The current paper extends our earlier investigation. The core objective is to boost detection accuracy, primarily by reducing false positives and thereby improving precision. We introduce an innovative technique that supplies the multimodel neural network with a series of successive images rather than relying on a lone fixed image. Because overlapping occurs among the following images, the identical target must be recognized in several frames. Correlation across these adjacent images was determined, with identified zones from one frame shifted onto others using the displacement vector. The key idea is that a target spotted across multiple frames is more likely to be a genuine detection, since the model is unlikely to flag the same false positives repeatedly. Informed by this observation, three distinct algorithms are outlined for eliminating detections and propagating detections across images within the sequence. Each one achieved a precision of roughly 80%, marking a nearly 20% increase over existing leading methods.

Keywords Convolutional neural networks, Aerial images, Displacement vector, Search and rescue

*Correspondence:
Maria Silva
maria.silva@uc.pt
[1] Department of AI for Cork Oak Acorn Yield Forecasting, Faculty of Agriculture, University of Coimbra, Coimbra, Portugal
[2] Department of Smart Forest Regeneration Monitoring, Faculty of Agriculture, University of Porto, Porto, Portugal
[3] Department of Machine Learning for Wildfire Fuel Load Prediction, Faculty of Agriculture, University of Lisbon, Lisbon, Portugal

Introduction

A search and rescue (SAR) mission seeks to locate a missing individual while still alive and deliver urgent support without delay. Consequently, timing represents a decisive element. Shortening the overall search period can be greatly facilitated by Unmanned Aerial Vehicles (UAVs or drones) that automatically and consistently capture the target zone. Maximizing the value of drones here requires skillful handling of sensor data. This creates demand for a system that automatically identifies people in drone-captured images during SAR activities.

Hence, the primary purpose of this study is to build a detection model for individuals in overhead imagery, suitable as an aid for live SAR missions and outperforming existing or suggested alternatives. The work outlined here builds on the earlier study by Kundid Vasić and Papić [1], which processes fixed aerial views along with contextual details and feeds them into a multimodel neural network designed to detect people in nonurban landscapes recorded by UAVs. That earlier system reached a recall of 94.66% and a precision of 68.90%. Still, these figures leave room for further gains, above all in precision. Reduced precision often stems from elevated false-positive counts. Our continued efforts, therefore, aim to reduce false positives and improve precision. To this end, we introduce a method that uses image sequences as input to

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium or format, provided that appropriate credit is given to the original author(s) and source, a link to the license is provided, and any changes made are indicated. Adapted material must be distributed under the same license.

© 2025 The Author(s). | Published by Agro Intelligence Research Press

the system. This method rests on the observation that targets appearing in several successive frames are more likely to be authentic detections. In contrast, those visible in just one frame tend to be erroneous. Three distinct algorithm categories grounded in sequence-based person detection are proposed, each delivering stronger performance than the approach reported by Kundid Vasić and Papić [1].

Training and evaluation of the model occurred on the HERIDAL database (http://ipsar.fesb.unist.hr/HERIDAL%20database.html, accessed on 18 November 2021), developed specifically for identifying humans and similar objects from the air and detailed in Božić-Štulić *et al.* [2]. The HERIDAL collection holds 1684 drone photographs of nonurban zones taken under simulated SAR conditions across multiple sites. Every image features at least one person in a variety of positions. Model training used 1583 images, with 101 reserved for testing. Since the HERIDAL collection lacks image sequences vital to this research phase, an additional group of aerial photographs replicating SAR conditions was assembled. These fresh photographs were arranged into sequences and employed to evaluate the baseline model. No retraining took place, so gains in detection quality arise solely from the newly introduced algorithms described here. Our key contributions include:

- A novel, enhanced model for identifying persons in aerial image sequences tailored to SAR situations
- A system leveraging positional data of objects found in adjacent images to preserve or remove candidate regions
- Achievement of better outcomes than the present top-performing techniques for person spotting in aerial views

The paper proceeds as follows. Section 2 covers relevant prior studies on this topic. Section 3 outlines the current leading model for person detection in wild-area aerial images: RFCCD—RPN+FPN+Classification+Context +Deviation [1]. Its performance when applied to the HERIDAL database and corresponding outcomes is also discussed. This serves as the foundation for the present work. Accordingly, Section 4 examines results from applying this model to a brand-new collection of image sequences assembled specifically for testing in this study. The subsequent section details an algorithm that refines RFCCD outcomes using displacement vector data from neighboring images. Section 6 reports the findings from the proposed algorithms, with full explanations. A final section summarizes the conclusions.

Related Work

SAR missions represent organized efforts to locate and provide proper aid to missing individuals. The process involves four core phases: locate, access, stabilize, and transport the missing person [3]. SAR activities fall into several types based chiefly on landscape or environment, generally grouped as urban search and rescue (USAR) or wilderness search and rescue (WiSAR). USAR addresses extraction from ruined buildings or city infrastructure, where victims are commonly trapped beneath debris. Triggers range from natural events (earthquakes, floods, tornadoes, hurricanes) to human-caused incidents (war, terrorist attacks). WiSAR, by contrast, covers searches in open countryside (caves, sea, mountains, plains, etc.). This study addresses wilderness SAR.

Common to every SAR effort are demands for rapid and effective response to limit further harm or loss of life, given that delays can produce immediate and grave outcomes for those affected [4]. Frequently, rescue efforts are slowed by challenges in determining the missing person's location, which means the initial phase (locating the individual) consumes critical time.

Given the importance of time, ongoing innovation in search-acceleration tools remains essential. In response, Unmanned Aerial Vehicles (UAVs) have become a fundamental strategic asset in SAR activities [5, 6]. UAVs provide high-quality overhead imagery promptly and reliably, enabling faster decision-making and providing meaningful support to SAR personnel [7].

© 2025 The Author(s). | Published by Agro Intelligence Research Press

A substantial portion of research on UAV deployment in SAR relies on thermal imaging to detect humans by exploiting differences in heat from the environment [8-10]— Rudol and Doherty [10] combined thermal and colour cameras for onboard UAV detection. Infrared frames were first reviewed to outline human shapes, followed by colour data for body classification. Imagery from SAR flights can be analyzed directly on the UAV or transmitted elsewhere for processing. Yet analyzing detailed, high-resolution frames requires significant computational power, which strains UAV hardware. During transfer to ground control, images must avoid compression to prevent data degradation that could complicate the location of tiny targets. While lossless compression options exist, they still impose heavy processing loads on the UAV for large images, proving often unworkable and time-consuming. Thus, forwarding uncompressed original images from UAVs to ground stations for later analysis is more suitable. A capable high-resolution image transmission framework appears in Papić *et al.* [11].

The challenge of detecting persons in aerial images captured over nonurban landscapes can be viewed as a classic small-object detection problem, particularly at high altitudes. Operating at higher altitudes provides a broader field of view, thereby reducing the time needed to cover the full search region. Such images feature highly intricate backgrounds, with each person occupying less than 0.1% of the total image area, thus demanding high-resolution captures. In this scenario, the images measure 4000 × 3000 px. Generally, “small objects” are defined as those occupying an extremely small number of pixels within the image (less than 1% of the total area). Over recent years, small object detection has attracted considerable interest among computer vision researchers, largely due to its extensive real-world applications, such as autonomous driving [12], traffic monitoring [13], landfill monitoring [14], robotics [15], and video surveillance [16].

Nevertheless, object detection in aerial imagery goes beyond typical small-object detection challenges. It presents significant difficulties not only due to the tiny size of the targets but also owing to the dense presence of irrelevant elements (such as stones and vegetation), fluctuations in viewpoint and illumination conditions, and frequent partial or total occlusion of the relevant objects. By reviewing the historical progression of object detection techniques [17], it becomes evident that a key breakthrough occurred alongside the fast advancement of deep convolutional neural networks. Traditional non-neural methods first involve explicit feature extraction (for example, edges, corners, and shapes) followed by separate classification steps. In contrast, neural network-based methods enable fully end-to-end detection without manual feature engineering. Consequently, the latter are frequently used for small-object detection tasks [18-21].

Most deep learning approaches to object detection rely on the core components of Convolutional Neural Networks (CNNs), specifically the convolutional and pooling layers. A convolutional layer produces a feature map that precisely encodes the location of detected features from the input. As a result, even slight alterations to a feature in the source image (such as rotation or cropping) will lead to a distinct feature map. Pooling layers promote generalization by condensing information from specific regions of a feature map. The resulting new feature map retains essential characteristics while eliminating minor details. This process introduces down-sampling and decreases the overall number of parameters requiring training. Nonetheless, this mechanism can be detrimental to the detection of small objects. For instance, a 32 × 32-pixel object in the input image would shrink to a single pixel after five pooling layers with a stride of 2, often resulting in missed detections. To illustrate downsampling effects more clearly, a representative CNN architecture, VGG16 [22], is depicted in **Figure 1**. This design incorporates five pooling layers with a stride of 2, resulting in five-fold downsampling relative to the original input.

© 2025 The Author(s). | Published by Agro Intelligence Research Press

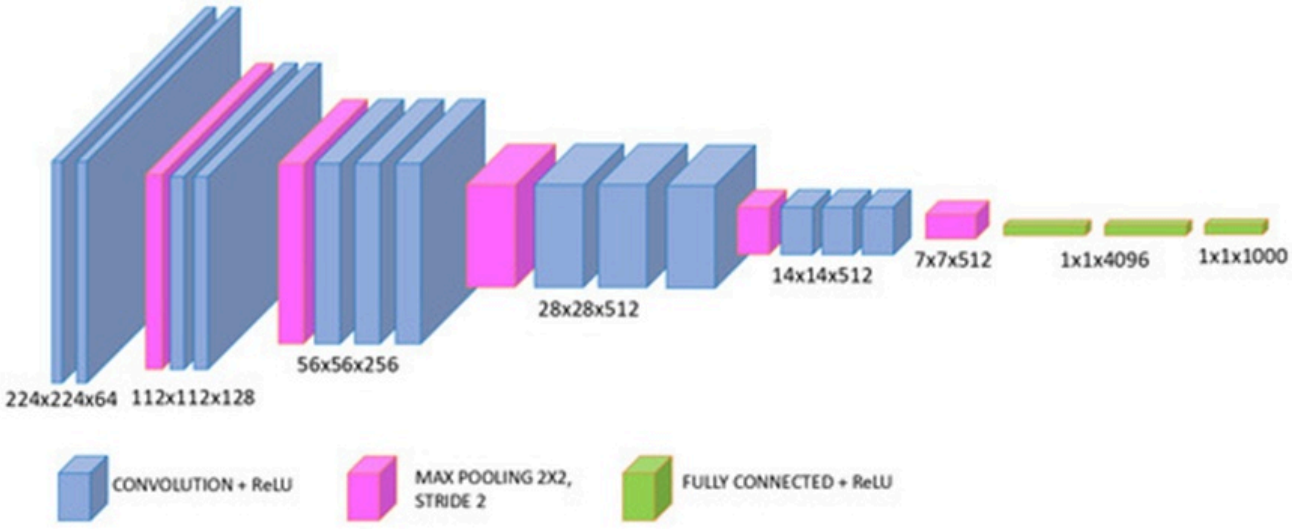

Figure 1. CNN architecture: VGG16.

This downsampling characteristic also explains why even advanced object detection frameworks, such as Fast Region-Based Convolutional Neural Network (R-CNN) [23] and Faster R-CNN [24], often struggle to detect small objects [25]. To overcome these limitations, conventional deep learning techniques are regularly adapted using multiple strategies to support effective small-object detection and deliver improved outcomes [26-28]. A detailed overview of recent deep learning techniques for small-object detection appears in Liu *et al.* [29]. Within the specific context of aerial image object detection, a major difficulty lies in identifying targets from varied aerial viewpoints and angles, across different poses, including cases of partial occlusion or movement. Humans can readily identify objects in images despite changes in rotation, scale, translation, or partial blocking.

In contrast, this process is considerably more difficult for computer vision systems due to their computational approach. Furthermore, human detection in aerial imagery often yields a substantial number of false positives. Given that images acquired during SAR missions form sequences, the persistence of a detected object across multiple frames can help filter out false positives [30, 31]. Analysis of image sequences is a common practice in computer vision, as sequences provide richer information than individual frames, particularly regarding motion dynamics. This proves especially valuable in areas such as video object detection [32], autonomous vehicles [33], object tracking [34, 35], and similar applications. That said, notable distinctions exist between aerial target tracking and conventional ground-based tracking, stemming from factors such as limited target pixel resolution, frequent occlusions, and low contrast between targets and backgrounds [36]. To minimize false alarms, some researchers have developed moving-object detectors that integrate spatial and temporal cues. In the study by Shen *et al.* [37], temporal saliency helps generate a rough segmentation, while spatial saliency refines object details within potential motion areas. Even so, this technique continues to suffer from elevated false-alarm rates in complex scenarios with cluttered backgrounds, occlusions, and varying illumination. The method described by LaLonde *et al.* [38] employs a two-stage process: the initial stage merges motion and appearance data inside a CNN to suggest regions of interest, which are then refined in the second stage for vehicle localization. However, this strategy performs poorly on smaller targets, as key visual attributes such as color, shape, and texture become difficult to discern.

RFCCD Model for Person Detection

Person detection in aerial images for SAR operations represents a highly specialized task, and suitable public datasets for this application remain scarce. As a result, Božić-Štulić *et al.* [2] created a dedicated dataset, HERIDAL, comprising aerial views of nonurban environments. To ensure the images realistically reflect actual SAR conditions, the creators drew upon statistical data and expert insights from SAR activities [39]. This same dataset was utilized in our earlier study,

© 2025 The Author(s). | Published by Agro Intelligence Research Press

whose primary objective was to design a multimodel deep learning solution for detecting humans in aerial imagery to assist SAR efforts. Various deep learning techniques were explored, leading to the introduction of an entirely novel method. This proposed solution incorporates several neural networks across both the region proposal and classification phases. The different variants included:

• RPNC (RPN+Classification)—a conventional RPN (Region Proposal Network) model [24] for generating region proposals, with the suggested regions subsequently classified by a custom neural network for binary classification trained and evaluated on image patches extracted from the HERIDAL dataset;

• FPNC (FPN+Classification)—an approach comparable to RPNC, except that FPN (Feature Pyramid Network) [40] is employed during the region proposal phase;

• RFC (RPN+FPN+Classification)– an innovative multimodel strategy that integrates both RPN and FPN architectures for region proposal generation;

• RFCC (RPN+FPN+Classification+Context)—a newly introduced multimodel technique similar to RFC, but enhanced in the classification phase by employing two separate neural networks to evaluate the proposed regions;

• RFCCD (RPN+FPN+Classification+Context+Deviation)—an advanced multimodel framework building upon RFCC by incorporating an additional step that discards regions exhibiting low standard deviation to decrease false positive detections.

Performance assessment of all approaches relied on standard precision and recall metrics. Precision quantifies the proportion of true positives among all detections (including both false and true positives). Recall measures the proportion of true positive detections relative to the total number of relevant objects indicated in the ground truth annotations, meaning all objects that ought to have been identified. The formulas for computing these metrics are provided in **Eq. 1**, where True Positive (TP) denotes correctly identified objects of interest, False Positive (FP) denotes incorrectly detected non-relevant objects, and False Negative (FN) denotes missed objects of interest.

$$Precision = \frac{TP}{TP + FP}; Recall = \frac{TP}{TP + FN} \quad (1)$$

Elevated recall values correspond to a greater count of TP detections (and fewer FN cases), whereas higher precision reflects a reduced number of FP detections. Given the nature of this study, which centers on locating a missing individual in an image—where any FN detection indicates failure to find the person—the total number of TP detections is paramount. Conversely, each FP detection misleads rescue teams to the wrong sites, resulting in wasted time and resources. Therefore, the primary objective is to attain a reasonable equilibrium between recall and precision. To assess this balance, the F-score metric is applied to indicate the system's accuracy and resilience. This metric is a combined value that represents the harmonic mean of precision and recall, as detailed in **Eq. 2**.

$$F = 2 * \frac{(precision * recall)}{(precision + recall)} \quad (2)$$

The conventional F-measure is typically referred to as the $F1$ measure, with the number 1 indicating balanced weighting between precision and recall. Nonetheless, it is feasible to emphasize one metric over the other through the broader $F\beta$ measure formula in **Eq. 3**, where the parameter β determines the relative weight given to recall versus precision.

© 2025 The Author(s). | Published by Agro Intelligence Research Press

$$F_\beta = (1 + \beta^2) * \frac{precision * recall}{(\beta^2 * precision) + recall} \quad (3)$$

When maximizing precision is prioritized, maintaining a smaller overall detection count is preferable to allowing excessive FP detections. In contrast, maximizing recall generates many TP detections irrespective of the accompanying FP count. Because TP detections directly indicate the location of individuals, recall is clearly more significant in this application, although quantifying the exact degree remains challenging. Accordingly, beyond the standard $F1$ measure, the $F2$, $F5$, and $F10$ metrics were also computed to report the outcomes (corresponding to scenarios in which recall is weighted 2, 5, or 10 times as heavily as precision). All experimental outcomes for the tested methods on the HERIDAL dataset images are summarized in **Table 1**, where GT stands for Ground Truth (the actual number of individuals present in the images that require detection).

Table 1. Results of the detection in a new set of images using different models are presented in Kundid Vasić and Papić [1].

Algorithm	*F10*	*F5*	*F2*	*F1*	Recall	Precision	FN	FP	TP	GT
RPNC	94.34%	91.00%	75.84%	57.91%	95.55%	41.55%	15	453	322	337
FPNC	86.54%	86.22%	84.49%	81.45%	86.65%	76.84%	45	88	292	337
RFC	94.87%	92.96%	83.46%	70.15%	95.55%	55.42%	15	259	322	337
RFCC	94.55%	93.40%	87.38%	78.05%	94.96%	66.25%	17	163	320	337
RFCCD	94.31%	93.32%	88.07%	79.75%	94.66%	68.90%	18	144	319	337

Reviewing the performance data shows that the RPNC approach delivered stronger recall scores but still produced a high volume of FP detections, resulting in notably weaker precision. In comparison, the FPNC approach yielded fewer FP detections, thereby improving precision. However, this advantage was offset by a drop in TP detections (and a corresponding rise in FN detections). Due to the particular requirements of this application—where locating a missing individual is essential, and any FN detection represents a total miss of the person—the total count of TP detections is of utmost importance. For that reason, it would be inaccurate to describe this approach as the top selection for practical SAR missions. At the same time, every FP detection sends search teams to the wrong areas, resulting in wasted time and personnel effort. Accordingly, the primary target is to achieve the highest possible TP detections and the lowest possible FP detections.

To merge the positive aspects of the RPNC approach, known for delivering extensive TP detections, together with the FPNC approach, recognized for limiting FP detections, a fresh multimodel solution called RFC was introduced. This developed model fully utilizes the RPNC framework by maintaining the same TP detection count as RPNC and, equally importantly, reducing FP detections by leveraging the FPNC framework. As the resulting recall score reached a sufficiently high level, continued efforts shifted toward further reduction of FP. In accordance, the RFCC framework was created, building on the RFC structure by adding contextual details in the classification phase. This addition employs pixel-based context [41]. The emerging findings confirm that integrating contextual characteristics from nearby zones in the classification phase led to a substantial drop in FP detections while preserving the peak TP detections provided by RPNC.

On top of that, an extra decrease in FP detections occurred through the elimination of detected regions displaying low pixel-level standard deviation (in the RFCCD model). Although TP detections experienced a small reduction, the

© 2025 The Author(s). | Published by Agro Intelligence Research Press

corresponding increase in precision enabled this framework to achieve the best overall equilibrium. It may therefore be stated that, among the available techniques, RFCCD is the most suitable for real-world use in SAR operations.

RFCCD architecture

A schematic representation of the developed multimodel system is presented in **Figure 2** and may be condensed into the following main stages:

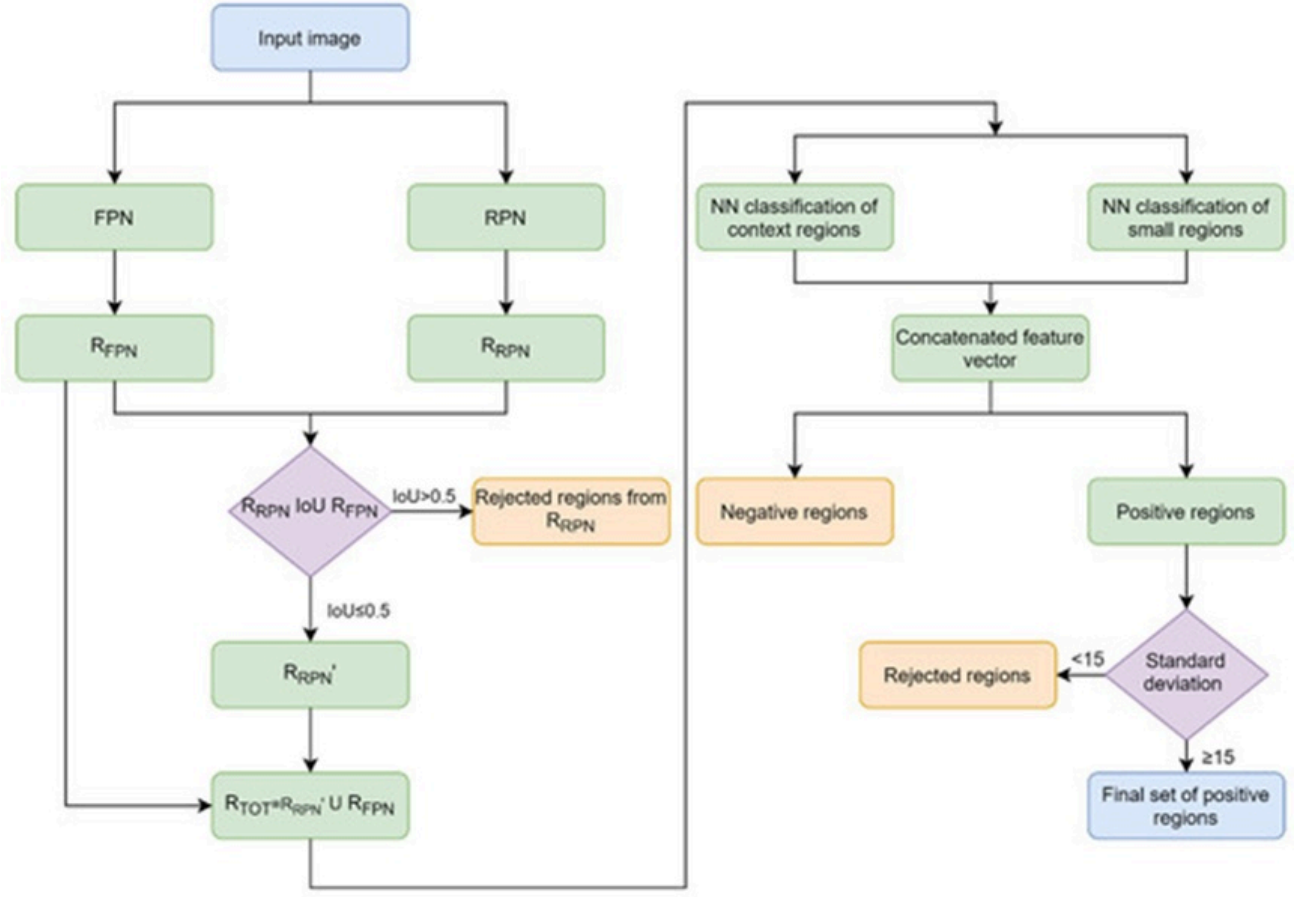

Figure 2. RFCCD architecture diagram.

· Input images undergo division into blocks measuring 500 × 500.

· RPN and FPN techniques process the image blocks to generate candidate regions of interest.

· Intersection over Union (IoU) metrics (the area of overlap between two bounding boxes divided by their union) are computed across regions from both the RPN and FPN model proposals. If the IoU exceeds 0.5, the overlapping regions are excluded from the RPN proposal collection because they resemble entries in the other collection. The purpose is to prevent the repeated identification of identical objects (whether true positives or false positives). The remaining regions from the RPN collection, along with all regions suggested by the FPN model, are merged to form a single consolidated collection for subsequent processing.

· After finalizing the collection of proposed regions, they proceed to the classification phase.

· Starting from the central pixel within each proposed region, an 81 × 81 square area is generated and submitted for classification using a custom-designed neural network specifically for this purpose. The initial layer consists of a

© 2025 The Author(s). | Published by Agro Intelligence Research Press

convolutional operation with 32 filters, followed by a pooling layer (3 × 3 with a stride of 3). A second convolutional layer with 32 filters, and finally another pooling layer (3 × 3 with a stride Next come two convolutional layers employing 64 filters before reaching a fully connected layer. This neural network was crafted empirically, trained, and validated using patches extracted from the HERIDAL dataset, ultimately achieving an accuracy of 99.21%.

· Concurrently, an area sized 243 × 243 (centered similarly on the proposed region's central pixel) gets formed, incorporating surrounding contextual details about the detected object's environment, and undergoes classification through the VGG16 network via transfer learning. Transfer learning is a method in which knowledge acquired during training on one task serves as an initial foundation for adapting the network to a different task. Here, weights originating from the VGG16 network pretrained on the ImageNet dataset [42] for person classification provided the foundation for retraining the network on our contextual region classification. This step enhanced the overall effectiveness of the classification phase.

· Both classification processes run only through the feature vector extraction phase, after which the feature vectors produced by each network are combined into a unified vector, followed by a fully connected layer and a softmax classification layer.

· As the final step, the regional standard deviation is computed for every region labeled as positive. Regions with a standard deviation below 15 are eliminated, while the remaining regions qualify as confirmed detected objects.

Deploying the RFCCD Model across Image Sequences

Evaluation of the RFCCD model is performed on a separate collection of image sequences. This step supports an accurate side-by-side assessment of its outcomes against the results produced by the additional models described in the sections ahead.

Acquisition of the testing data

Because the HERIDAL database contains no sequences captured repeatedly at the same sites, assembling new aerial photographs that reflect genuine SAR conditions is a vital step in this research phase. Notably, the RFCCD model outlined in Section 3 was not retrained on the newly collected images. The original version, developed through training neural networks solely on the HERIDAL database, remains in active use. To thoroughly test this model on unseen imagery, a dedicated new collection was assembled. It features sequences consisting of three successive images recorded over identical ground areas. Empirical testing established three as the ideal sequence length. Pairs of only two successive images were insufficient to verify detection reliability confidently. As an illustration, an object spotted in the opening image yet absent in the following one (or vice versa) leaves considerable doubt about whether the sighting qualifies as genuine. Hence, a minimum of three successive images becomes necessary. This setup permits any object appearing in at least two of the three frames to be registered as a true positive. Sightings limited to a solitary image are classified as false positives and are subsequently excluded from the final list of confirmed detections.

At the same time, sequences extending past three images prove equally unsuitable. Real SAR missions commonly involve expansive search zones that demand broad coverage, inevitably producing a large volume of images. Raising the capture rate enlarges the "overlap" zone between successive shots. When intervals remain fixed and allow one object to register across four successive frames, the adjacent-image overlap must reach at least 75%. Under these parameters, the overlap between the initial and fourth images shrinks to 25%. Still, the probability of any single object remaining visible through all four frames stays extremely low. This capture strategy would generate an excessively large overall image volume. Greater image counts increase processing time, yet rapid turnaround is a decisive element in SAR missions. With this in mind, the effort centers on trimming processing demands wherever feasible. These considerations led to the determination that sequences of exactly three consecutive images offer the optimal configuration for methods relying on cross-frame detection analysis.

© 2025 The Author(s). | Published by Agro Intelligence Research Press

As a result, the fresh image collection compiled during this study comprises 99 images in total, arranged into 33 sets of three consecutive images, each set drawn from a distinct site. All photographs originated from nonurban terrain during two distinct seasons (a portion recorded in summer, when green shades prevail thanks to trees and ground vegetation, and others in autumn, dominated by grey and brown shades) to enhance variety across the entire collection. The acquisition was made with a DJI Spark drone carrying a 12MP camera, operating in free-flight mode. This approach produced differing overlap amounts between consecutive frames within each set, along with correspondingly varied correlation and displacement vectors. Several examples drawn from these consecutive image sets appear in **Figure 3**.

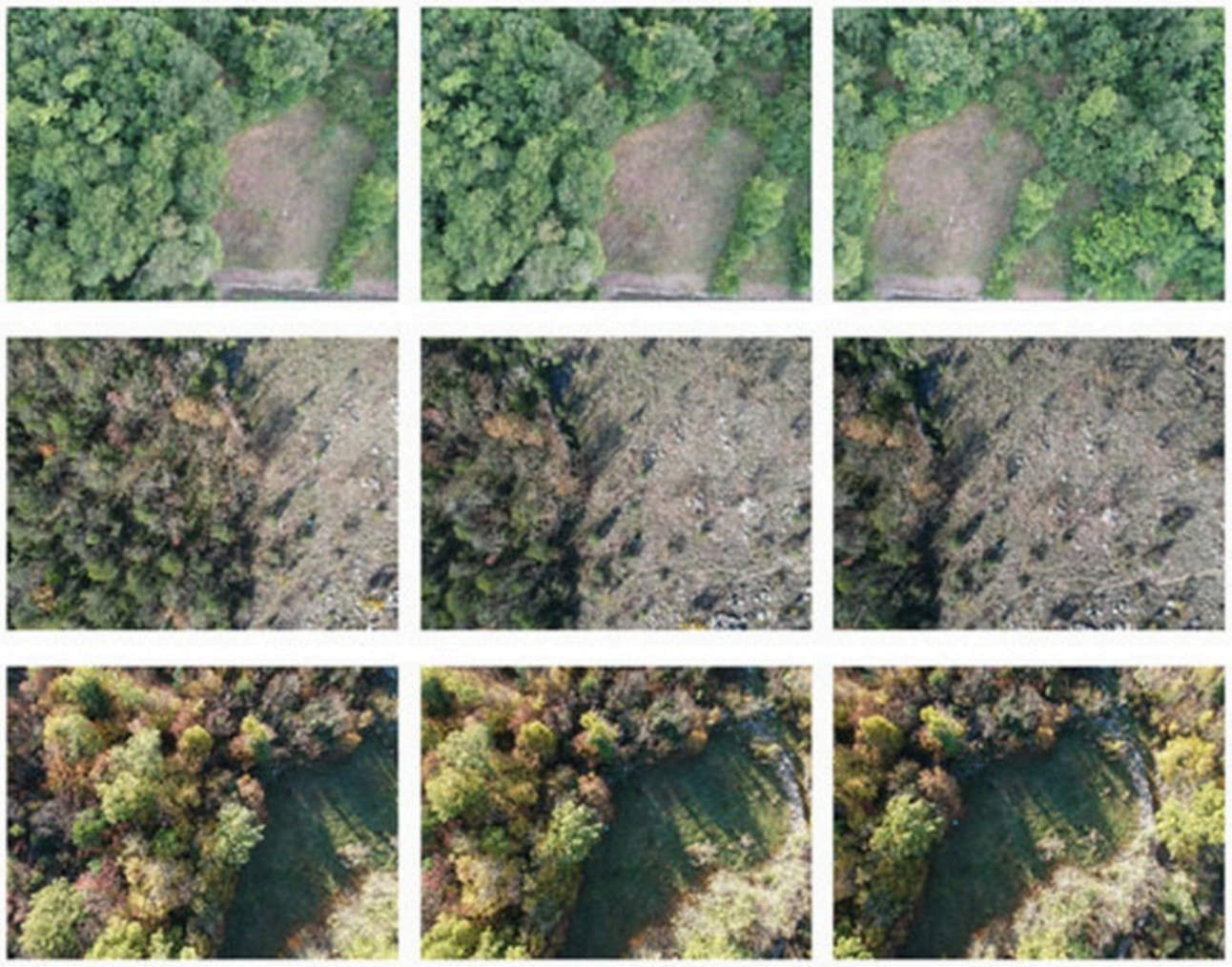

Figure 3. Examples of three image sequences.

Results of the RFCCD model in a new set of image sequences

Given that testing relied on a wholly fresh collection of images outside the HERIDAL database, a thorough assessment of the suggested RFCCD model became essential on this new collection. The obtained performance metrics are displayed in **Table 2**. It is worth noting that the model received no training whatsoever on these images; the version trained exclusively on the HERIDAL database was applied directly to test on the unfamiliar data. This application successfully identified 330 out of 339 individuals, yielding a recall of 97.35%. Precision is lower due to a higher false-positive rate, suggesting room for further refinement. The proposed strategy focuses on reducing false-positive detections and improving overall outcomes by using an algorithm that computes displacement vectors between successive images within each sequence group.

© 2025 The Author(s). | Published by Agro Intelligence Research Press

Table 2. Results of the detection in a new set of images using the RFCCD model.

Algorithm	GT	TP	FP	FN	Precision	Recall	$F1$	$F2$	$F5$	$F10$
RFCCD	339	330	176	9	65.22%	97.35%	78.11%	88.61%	95.54%	96.87%

Proposed Algorithms for Improving Results of the RFCCD Model

To elevate the outcomes delivered by the RFCCD model, this study introduces three distinct algorithmic variants. Each variant draws on the correlation between successive images. Consistent with the objects identified in a given frame and the derived correlation across images, the algorithm estimates the positions of those objects in subsequent frames using displacement vector estimation. Hence, computing the correlation forms the opening phase.

Once the correlation computation finishes, the detected regions from one image are shifted to the two successive images according to the displacement vector. Next, the IoU is calculated between the genuine detections produced by the RFCCD model in one image and the translated detections derived from another image. If the IoU exceeds 0.5, the regions are considered overlapping, indicating that the same object has been identified in both images. Furthermore, to account for possible inaccuracies in image correlation, a tolerance buffer was implemented. This effectively widens the zone where detected objects across successive frames are deemed identical. Here, the tolerance value was set empirically at 200 px.

The core assumption is that an object identified in at least two of the three successive images is most likely a true positive, given the low probability that the detection model would locate the same false positive across separate images. Drawing on this assumption, three algorithmic variants were developed to refine the RFCCD model's outcomes. The algorithms vary according to their methods of rejecting or incorporating detected objects, as specified below:

• RFCCD + Displacement Vector (RFCCD+DV)—removing detected regions that emerge in merely one of the three successive images;

• RFCCD + Displacement Vector and Adding (RFCCD+DVA)—removing detected regions that emerge in merely one of the three successive images while incorporating detected regions that surface in two successive images into the third successive image;

• RFCCD + Displacement Vector and Adding with Classification (RFCCD+DVAC)—removing detected regions that emerge in merely one of the three successive images while incorporating detected regions that surface in two successive images into the third successive image, but exclusively when the detected object lies inside an area covered by trees or forest shrubs.

The suggested methods are described in more detail in the subsequent sections.

Calculating the correlation between consecutive images

If $S=[I1,I2,I3]$ signifies a collection of successive images, and $Ii,i=[1,\ldots,3]$ signifies the images contained within that collection, the algorithm processes pairs of images $ccor(In,Im),n\neq m,n\in i,m\in i$. It derives the displacement vector expressed through the distance and direction of movement. In the initial phase of this algorithm, the higher spectral image undergoes conversion into a two-dimensional matrix $Ii=M[r,c]$, where the three-dimensional vector $pk,j\in Ii,pk,j=[rk,j,gk,j,bk,j]$ encodes light intensity across a given spectrum using RGB components rk,j,gk,j,bk,j within the interval $[0,\ldots,255]\in\mathbb{N}$, transforming it into the one-dimensional vector $gk,j\in Gi,gk,j=[ek,j],ek,j\in\mathbb{R}$. The procedure for converting a multi-spectral image into an intensity image is presented in **Eq. 4**.

© 2025 The Author(s). | Published by Agro Intelligence Research Press

$$e_{k,j} = \frac{r_{k,j} + g_{k,j} + b_{k,j}}{\sum_{k=0}^{r} \sum_{j=0}^{c} p_{i,j} * \frac{1}{r*c}} \quad (4)$$

A convolution operator gets applied to both images (Gn,). $\mathscr{F}$ serves as the operator for the Fourier transform, and in accordance with the convolution theorem, **Eq. 5** applies.

$$G_n * G_m = F^{-1}\{F\{G_n\} * F\{G_m\}\} \quad (5)$$

To acquire a similarity metric between the two images, a cross-correlation calculation is required. Given that cross-correlation corresponds to a convolution in which the second signal is horizontally and vertically mirrored, the second image was rotated by 180∘. To establish the distance between the two images in terms of displacement D and the angle of displacement θ, the auto-correlation or self-similarity $Gn*(Gn)$ must first be calculated, followed by the cross-correlation between the two images $Gn*M(Gm)$. An illustration of the outcomes for these two metrics is shown in **Figure 4**, along with the images used for their computation. An example of the auto-correlation appears above, while the cross-correlation measure between the two images appears below.

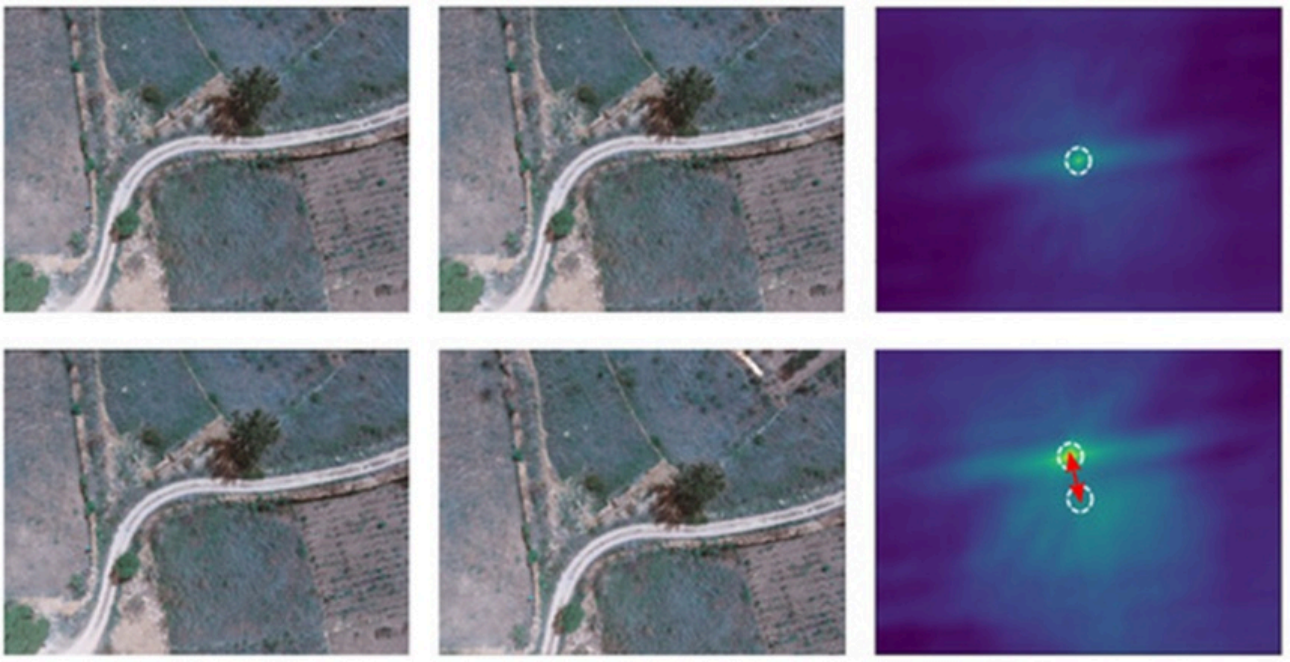

Figure 4. Example of auto-correlation (above), cross-correlation (below), and displacement D (red arrow).

The opening action in deriving the displacement D and the angle of displacement θ between the two images $Gn*(Gn)$ and $Gn*M(Gm)$ consists of identifying the point exhibiting the highest intensity. This point is determined by computing the Euclidean distance using the maximum argument function on the auto-correlation function $L1=ar(Gn*M(Gn))$, where $L1=(Lx1,Ly1)$ designates the point of maximum intensity. The same procedure applies for the cross-correlation $L2=ar(Gn*M(Gm))$. The computation of the distance separating points $L1$ and $L2$, representing the displacement D and the angle of displacement θ, is illustrated in **Eqs. 6 and 7**. This derivation of D and θ occurs for every pair of images within the image collection S.

© 2025 The Author(s). | Published by Agro Intelligence Research Press

$$D = \sqrt{\left(L_{x_2} - L_{x_1}\right)^2 + \left(L_{y_2 - L_{y_1}}\right)^2} \quad (6)$$

$$\theta = tan^{-1}\left(L_{x_2 - L_{x_1}}, L_{y_2 - L_{y_1}}\right) - \frac{\Pi}{2} \quad (7)$$

Based on these parameters, the detected regions were translated from one image to another. The detected regions correspond to regions of interest defined as $ROI(Ii)=[roi1,\ldots,roiw]$,$roiu=[xu,yu,wu,hu]$,$u\in[1,\ldots,s]\subset\mathbb{N}$, where s denotes the total quantity of proposed regions in one image. Following this, verification can occur regarding which regions of interest correspond in images $I1$ and $I2$ by computing the cross-section among all pairs of regions of interest. The operator for confirming the cross section is defined by a function that shifts all regions of interest from one image to another using the displacement vector, as shown in **Eq. 8**.

$$tran\,(roi_u) = [x_u + D_1 * sin\theta_1, y_u + D_1 * cos\theta_1, w_u, h_u] \quad (8)$$

It is likewise essential to establish the distance between the two regions via the Euclidean distance from the upper left corner point of the region of interest, as shown in **Eq. 9**.

$$dist\,(roi_1, roi_2) = \sqrt{(x_2 - x_1)^2 - (y_2 - y_1)^2} \quad (9)$$

Applying this function across all regions from the set of ($I1$) generates an approximation of the regions of interest anticipated to exist in figure $RO(I2)$. A central function for detecting overlaps between regions is the cross-section operator in **Eq. 10**.

$$ROI\,(I_1) \cap ROI\,(I_2) = \begin{cases} roi_{1_u} & if \quad dist\left(tran\,(roi_{1_u}), roi_{2_k}\right) < 200 \\ \varnothing & else \end{cases} \quad (10)$$

Using correlation for estimating detection accuracy

Relying on the idea that any detection appearing in just one out of three successive images is very likely a false positive, the initial solution focused on removing such detections from the final results. This method is named “RFCCD+DV” (RFCCD + Displacement Vector). When applied to the objects identified by the RFCCD model, the algorithm performs correlation analysis between successive images, as explained in Section 5.1. Correlation is computed for every pair within a set (first-second, second-third, and first-third). Detected objects from each image are then shifted to the other images, and overlap — including the defined tolerance — is checked between the shifted detections and the actual detections present in the target image. If a matching overlap is found in at least two of the three images, the detections are treated as the same object and kept. All detections limited to a single image are discarded. This approach is the most direct and effective option, as it sharply cuts down false positives while preserving the original level of true positives. These gains are visible in the outcomes discussed in Section 6.

© 2025 The Author(s). | Published by Agro Intelligence Research Press

During testing of this algorithm, it was observed that some objects appearing in two of the three successive images were actual persons missed by the model in the remaining image. To address this, a new variant, RFCCD + Displacement Vector + Adding (RFCCD+DVA), was created. It operates like the previous method but adds one key step. Since objects seen in two images are assumed to be true positives, they can be inserted into the missing third image to raise the total true positive count and improve recall. Displacement vector information guides this insertion. Using the vector values calculated between the first-second or second-third pairs, the detection is projected to its expected spot in the third image and added to the results. This variant boosts recall yet lowers precision, according to the findings in Section 6. The precision drop occurs because many added detections are false positives.

It is evident that not all objects detected in two images are genuine, so including them in the third image also increases the false-positive rate. Researchers believed further gains were possible by factoring in the type of surroundings around each detection. This led to the development of RFCCD + Displacement Vector + Adding with Classification (RFCCD+DVAC). The core concept is to classify the environment of each detected object. A person in open ground is usually reliably captured by the RFCCD model. Still, in wooded areas, the individual may be hidden by trees and remain undetected in a single image of the sequence. Therefore, determining whether the surroundings are forested or open becomes necessary. The decision to insert a detection into the third image depends on this classification. Detections from the other two images are only transferred if the environment is identified as forest. This requires building an accurate classification model and selecting a suitable neural network architecture.

Speed and simplicity are critical to the model, as overly complex designs would slow the entire system and make it impractical for real use. For this reason, the task was simplified to binary classification with two categories: “forest” and “other”. The “forest” category covers areas with dense trees or similar features (such as heavy shadows) that could conceal a person. At the same time, “other” includes open spaces such as low vegetation, fields, paths, and similar terrain that offer little concealment.

Classification of the detected object environment

The neural network previously used successfully for binary classification tasks is detailed in Section 3.1, Step (5). Thanks to its strong performance and lightweight structure, it was reused for this stage of the system. First, a dedicated dataset had to be prepared for training and evaluation. Aerial images were split into 300 × 300 px blocks, from which 4857 blocks were selected. These were separated into training and validation groups. The training group had 1886 blocks labeled “forest” and 2028 labeled “other”. The validation group contained 472 “forest” blocks and 471 “other” blocks. The network is fairly shallow and processes 300×300 input blocks. It uses convolutional filters combined with ReLu nonlinear activation [43]. The output stage features two fully connected layers: one with ReLu activation and the other with sigmoid activation. Training ran for 50 epochs using a batch size of 32. The binary cross-entropy loss function [44] guided the process. Once trained, the model was tested on the validation blocks, reaching 82% precision and 79% recall. The moderate scores result from the fuzzy boundary between the two classes. Regions with sparse vegetation often resemble those with denser growth, making accurate separation difficult even for human observers.

This environment classification model was then integrated into the process of adding detections to the third image. Whenever the RFCCD model finds the same object in two successive images, the detection is transferred to the third image only when the surroundings are labeled “forest”. To make this decision, eight surrounding regions around the detected area are evaluated. The process starts from the central pixel of the detected region (also 300 × 300 px). This center point shifts 450 pixels left and 450 pixels up to establish the initial environment window. A sliding 3 × 3 grid generates nine regions total. The central region is skipped (as it matches the detected object), leaving eight regions for classification. If any of these eight is categorized as “forest”, the detection is added to the third image. This accounts for the chance that the person is present but concealed by trees. Although the environment classifier is less accurate than the person detector, experiments showed that the overall approach still delivers better results. Compared to simply adding every missing detection regardless of surroundings, this version significantly reduces false positives — achieving the main target of the improvement.

© 2025 The Author(s). | Published by Agro Intelligence Research Press

Results Obtained with the Proposed Algorithms

Table 3 displays the outcomes produced by the developed algorithms that utilize displacement vectors derived from image sequences. The table also includes the baseline performance of the RFCCD model on the same image collection for direct comparison. The RFCCD model produces a substantial number of false positives, highlighting the need for further enhancements. To achieve this, three separate algorithms relying on displacement vectors across successive images were introduced. The experimental findings confirm that each of the three algorithms outperforms the original RFCCD method. The initial algorithm, named “RFCCD+DV”, eliminates any detection that fails to register in at least two out of the three successive images. Relative to RFCCD, this version delivered a substantial boost in precision with no reduction in recall. Consequently, the F-measures also registered clear gains. Effectively, the algorithm removes all false positive detections confined to a single image. As objects appearing in at least two images are viewed as true positives, recall could be strengthened even more by incorporating those detections into the absent third image. This prompted the creation of the second algorithm, RFCCD+DVA, which discards single-image detections and transfers detections from two images to the third. An illustration of a single added true positive detection is shown in **Figure 5**. Every image within this sequence features three persons. The RFCCD model correctly located two persons across all three frames, but missed one person in the first image while detecting it in the second and third. The RFCCD+DVA algorithm successfully placed the missing true positive in the first image.

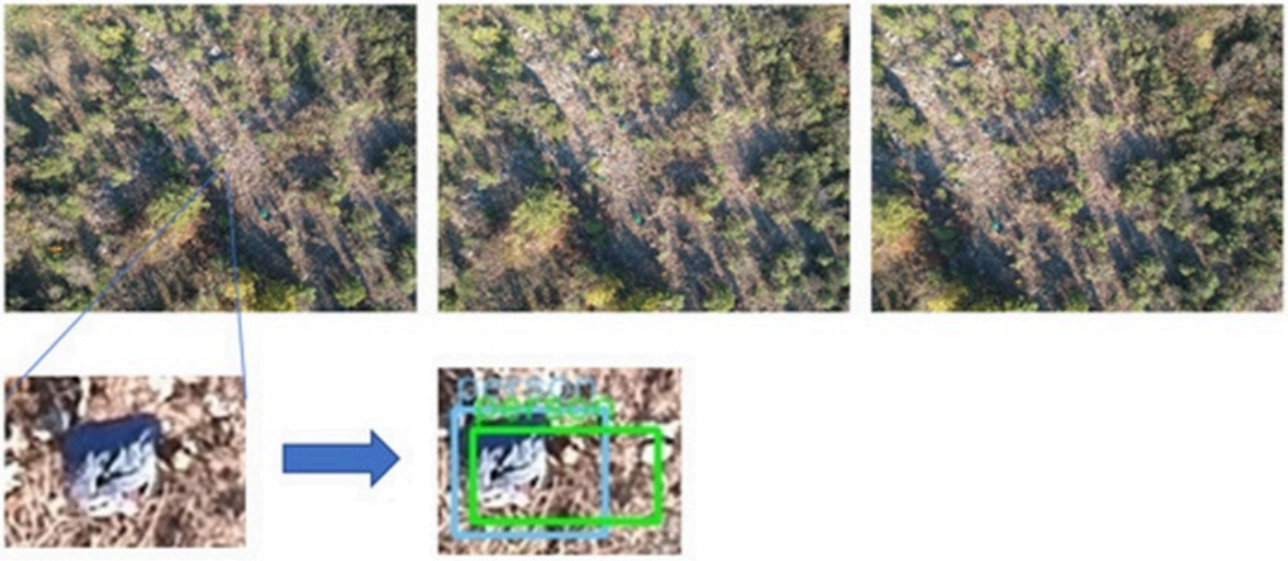

Figure 5. Example of a set of three images with one missed detection in the first image using the RFCCD model (above), an enlarged missed detection with the RFCCD model (below, left), and an enlarged added TP detection in this image using the RFCCD+DVA model (below, right).

Table 3. Results obtained with algorithms based on the displacement vector.

Algorithm	*F10*	*F5*	*F2*	*F1*	Recall	Precision	FN	FP	TP	GT
RFCCD	96.87%	95.54%	88.61%	78.11%	97.35%	65.22%	9	176	330	339
RFCCD+DV	97.20%	96.80%	94.56%	90.66%	97.35%	84.83%	9	59	330	339
RFCCD+DVA	97.99%	97.31%	93.64%	87.52%	98.23%	78.91%	6	89	333	339
RFCCD+DVAC	97.75%	97.23%	94.37%	89.49%	97.94%	82.38%	7	71	332	339

© 2025 The Author(s). | Published by Agro Intelligence Research Press

This algorithm improved recall but reduced precision, introducing 30 false-positive detections. As a result, all F-measures except F10 experienced a decline. To counter the addition of false positives, the third algorithm — RFCCD+DVAC — was formulated. Its approach restricts additions to the third image to detections whose local surroundings include forest, large trees, or similar features, since such settings increase the likelihood of temporary obstructions across the image sequence. Accordingly, the surroundings of each detection undergo classification, with only forest-classified cases transferred to the third image. This algorithm achieved superior precision compared to RFCCD+DVA, although recall dipped slightly (by 1 additional missed true positive).

Examination of the results indicates that the first method, RFCCD+DV, maintains the same number of true positive detections as the baseline RFCCD, thereby holding recall steady. At the same time, it reduces false positives by 65% and increases precision by nearly 20%. The surviving false positives that persist after filtering appear in at least two successive images and carry strong confidence scores. These typically involve objects that mimic human appearance from high altitude. Differentiating some of them from genuine persons can prove difficult even with direct visual review. Samples of such false positive detections are illustrated in **Figure 6**. Separately, **Figure 7** displays all 9 false negative detections. Interestingly, several false positives exhibit closer resemblance to human shapes than certain missed persons. This phenomenon arises from differences in occlusion levels, viewing perspectives, object shapes, and lighting variations.



Figure 6. Examples of FP detections (rocks, birds, shadows, bags, etc., detected as a person).



Figure 7. All FN detections (persons that are not detected).

Moreover, this algorithm delivers F-measure values that are equal to or only marginally below those of the other developed methods. In terms of F-score behavior, **Figure 8** demonstrates that all three algorithms produce superior F-score balances (across every β setting) than the original RFCCD model. The RFCCD+DV algorithm further stands out for its most consistent upward progression in F-score values as β varies. Given the challenge of selecting an optimal β, this consistent pattern offers a notable edge by showing limited sensitivity to the specific β chosen (maintaining high F-scores regardless). An additional strength is the algorithm's operational simplicity, which bypasses the need for supplementary review of detected objects or their surroundings. This results in reduced processing duration — a critical advantage in time-sensitive life-saving scenarios [4]. Based on these considerations, the RFCCD+DV algorithm represents the most effective choice for SAR missions.

© 2025 The Author(s). | Published by Agro Intelligence Research Press

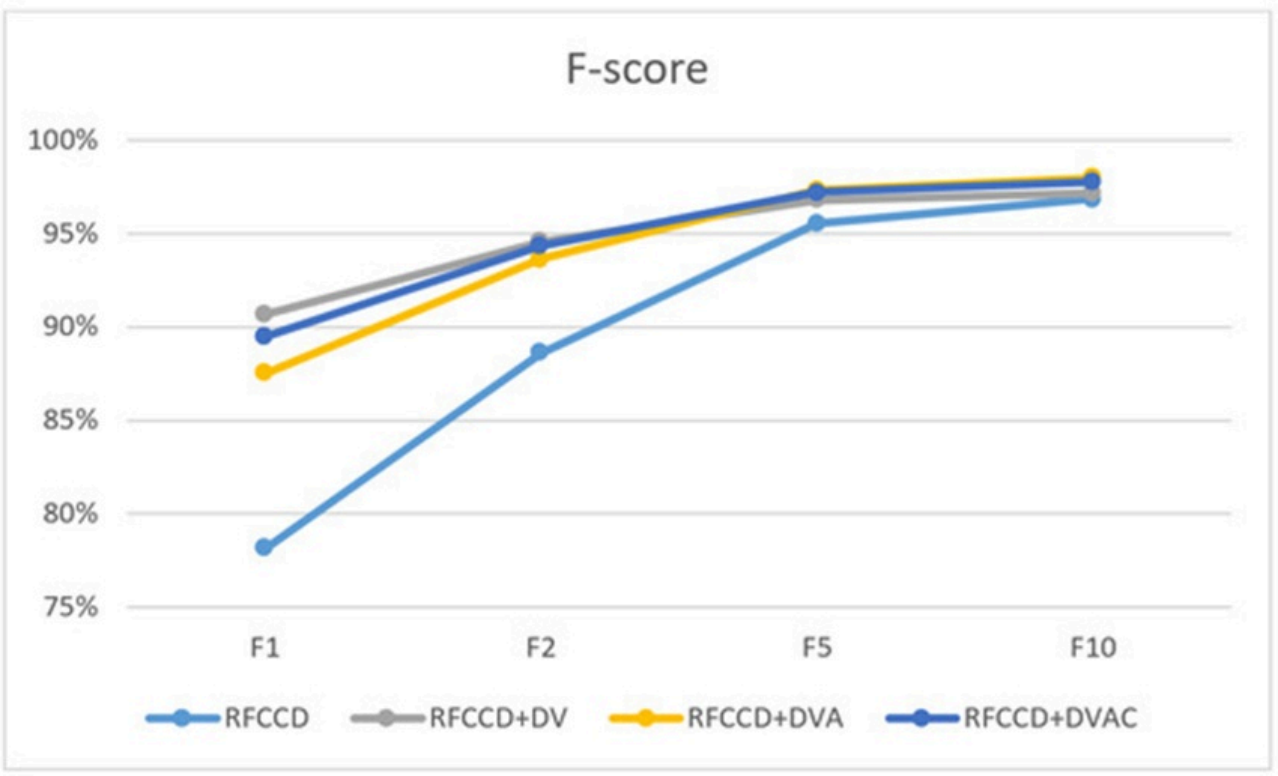

Figure 8. The value of F-score depends on the parameter β.

Conclusion

This study presents a collection of algorithms to refine person detection results in aerial image sequences intended for SAR missions. Every algorithm leverages displacement vector information to reduce false-positive detections. In practical SAR missions, false positives mislead rescue teams to incorrect areas and squander precious time. The proposed solution serves as a support tool, meaning the final processed images would still require manual visual verification to confirm potential positions of missing individuals. High volumes of false positives in an image would extend the time required for manual review, as each candidate must be examined. These high-resolution aerial captures (4000 × 3000) contain complex scenes and numerous potential false positives (including rocks, shadows, birds, and similar items that appear person-like from elevated viewpoints). The RFCCD model achieved reasonably solid performance, registering 176 false positives across 99 images (about 1.7 per image). This already permits fairly swift manual inspection. Nevertheless, real SAR missions typically generate extensive image volumes due to large search areas. Hence, the priority remains minimizing false positives per image to the greatest extent to preserve valuable time.

The algorithms detailed in this paper successfully lower false positives to below one per image on average while safeguarding or boosting true positive detections. Even though all three algorithms improve precision and maintain or increase recall, RFCCD+DV is considered the most suitable overall. This preference is based on two primary factors: it delivers the largest reduction in false positives and requires the least computational effort. Its simplicity is most apparent in the minimal additional processing time per image—an essential requirement for swift operations. Compared with alternative algorithms, RFCCD+DV imposes the smallest computational burden. Therefore, it shows strong potential as an auxiliary technique in actual SAR activities. Owing to its structural complexity, the approach is not designed for real-time execution directly on UAV platforms. Since locating the missing person is the paramount objective in SAR contexts, this research focused on achieving the highest possible detection accuracy. However, minimizing overall processing time remains equally important and will guide future efforts to optimize speed.

© 2025 The Author(s). | Published by Agro Intelligence Research Press

Acknowledgements

None

Conflict of interest

None

Financial support

This research was supported by the project “Prototype of an Intelligent System for Search and Rescue”, Grant Number KK.01.2.1.01.0075, funded by the European Regional Development Fund.

Ethics statement

None

Received: 06 Jan 2024 Revised: 12 Mar 2024 Accepted: 28 Mar 2024
Published online: 25 November 2025

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

References

1. Kundid Vasić M, Papić V. Multimodel deep learning for person detection in aerial images. Electronics. 2020;9:1459.
2. Božić-Štulić D, Marušić Z, Gotovac S. Deep learning approach in aerial imagery for supporting land search and rescue missions. Int J Comput Vis. 2019;127:1256-78.
3. Auerbach P. Wilderness medicine e-book: Expert consult premium edition—enhanced online features. Amsterdam (The Netherlands): Elsevier Health Sciences; 2011.
4. Adams AL, Schmidt TA, Newgard CD, Federiuk CS, Christie M, Scorvo S, et al. Search is a time-critical event: When search and rescue missions may become futile. WEM. 2007;18:95-101.
5. Waharte S, Trigoni N. Supporting search and rescue operations with UAVs. In: Proceedings of the International Conference on Emerging Security Technologies; Canterbury, UK; 2010. p. 142-7.
6. Karamanou A, Dreliosi GC, Papadimatos D, Hahlakis A. Supporting search and rescue operations with UAVs. In: Proceedings of the 5th International Conference on Civil Protection and New Technology; Kozani, Greece; 2018.
7. Półka M, Ptak S, Kuziora Ł. The use of UAV's for search and rescue operations. Procedia Eng. 2017;192:748-52.

8. Burke C, McWhirter PR, Veitch-Michaelis J, McAree O, Pointon HA, Wich S, et al. Requirements and limitations of thermal drones for effective search and rescue in marine and coastal areas. Drones. 2019;3:78.

9. Leira FS, Johansen TA, Fossen TI. Automatic detection, classification and tracking of objects in the ocean surface from UAVs using a thermal camera. In: Proceedings of the 2015 IEEE Aerospace Conference; Big Sky, MT, USA; 2015. p. 1-10.

10. 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; 2008. p. 1-8.

11. Papić V, Šolić P, Milan A, Gotovac S, Polić M. High-resolution image transmission from UAV to ground station for search and rescue missions planning. Appl Sci. 2021;11:2105.

12. Benjumea A, Teeti I, Cuzzolin F, Bradley A. YOLO-Z: Improving small object detection in YOLOv5 for autonomous vehicles. In: Proceedings of the International Conference on Computer Vision (ICCV 2021): The ROAD Challenge Workshop; Virtual; 2021.

13. Khan NA, Jhanjhi N, Brohi SN, Usmani RSA, Nayyar A. Smart traffic monitoring system using unmanned aerial vehicles (UAVs). Comput Commun. 2020;157:434-43.

14. Filkin T, Sliusar N, Ritzkowski M, Huber-Humer M. Unmanned aerial vehicles for operational monitoring of landfills. Drones. 2021;5:125.

15. Mittal S, Karthik MS, Kumar S, Krishna KM. Small object discovery and recognition using actively guided robot. In: Proceedings of the 2014 22nd International Conference on Pattern Recognition; Stockholm, Sweden; 2014.

16. Saha S, Vasegaard AE, Nielsen I, Hapka A, Budzisz H. UAVs path planning under a bi-objective optimization framework for smart cities. Electronics. 2021;10:1193.

17. Zou Z, Shi Z, Guo Y, Ye J. Object detection in 20 years: A survey. arXiv. 2019;arXiv:1905.05055.

18. Bejiga MB, Zeggada A, Melgani F. Convolutional neural networks for near real-time object detection from UAV imagery in avalanche search and rescue operations. In: Proceedings of the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS); Beijing, China; 2016. pp. 693–6.

19. Liu M, Wang X, Zhou A, Fu X, Ma Y, Piao C. UAV-YOLO: Small object detection on unmanned aerial vehicle perspective. Sensors. 2020;20:2238.

20. Han S, Yoo J, Kwon S. Real-time vehicle-detection method in bird-view unmanned-aerial-vehicle imagery. Sensors. 2019;19:3958.

21. Liang X, Zhang J, Zhuo L, Li Y, Tian Q. Small object detection in unmanned aerial vehicle images using feature fusion and scaling-based single shot detector with spatial context analysis. TCSVT. 2020;30:1758–70.

22. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv. 2014;arXiv:1409.1556.

23. Girshick RB. Fast R-CNN. In: Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV); Santiago, Chile; 2015. p. 1440–8.

24. Ren S, He K, Girshick R, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017;39:1137–49.

25. Nguyen ND, Do T, Ngo TD, Le DD. An evaluation of deep learning methods for small object detection. JECE. 2020;2020:3189691.

26. Zhang H, Wu J, Liu Y, Yu J. VaryBlock: A novel approach for object detection in remote sensed images. Sensors. 2019;19:5284.

© 2025 The Author(s). | Published by Agro Intelligence Research Press

27. Zhang S, Wu R, Xu K, Wang J, Sun W. R-CNN-based ship detection from high resolution remote sensing imagery. Remote Sens. 2019;11:631.

28. Liu T, Fu HY, Wen Q, Zhang DK, Li LF. Extended faster R-CNN for long distance human detection: Finding pedestrians in UAV images. In: Proceedings of the 2018 IEEE International Conference on Consumer Electronics (ICCE); Las Vegas, NV, USA; 2018.

29. Liu Y, Sun P, Wergeles N, Shang Y. A survey and performance evaluation of deep learning methods for small object detection. Expert Syst Appl. 2021;172:114602.

30. Wang H, Peng J, Yue S. A feedback neural network for small target motion detection in cluttered backgrounds. In: Proceedings of the 27th International Conference on Artificial Neural Networks; Rhodes, Greece; 2018. p. 728–37.

31. Wu D, Zhang L, Lin L. Based on the moving average and target motion information for detection of weak small target. In: Proceedings of the 2018 International Conference on Intelligent Transportation, Big Data Smart City (ICITBS); Xiamen, China; 2018. p. 641–4.

32. Koh J, Kim J, Shin Y, Lee B, Yang S, Choi JW. Joint representation of temporal image sequences and object motion for video object detection. In: Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA); Xi'an, China; 2021. p. 13370–6.

33. Wöhler C, Anlauf J. Real-time object recognition on image sequences with the adaptable time delay neural network algorithm—Applications for autonomous vehicles. Image Vis Comput. 2001;19:593–618.

34. Tissainayagam P, Suter D. Object tracking in image sequences using point features. Pattern Recognit. 2005;38:105–13.

35. Li W, Powers D. Multiple object tracking using motion vectors from compressed video. In: Proceedings of the 2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA); Sydney, NSW, Australia; 2017. p. 1–5.

36. Jia J, Lai Z, Qian Y, Yao Z. Aerial video trackers review. Entropy. 2020;22:1358.

37. Shen H, Li S, Zhu C, Chang H, Zhang J. Moving object detection in aerial video based on spatiotemporal saliency. CJA. 2013;26:1211–7.

38. LaLonde R, Zhang D, Shah M. ClusterNet: Detecting small objects in large scenes by exploiting spatio-temporal information. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); Salt Lake City, UT, USA; 2018. p. 4003–12.

39. Koester R. Lost Person Behavior: A Search and Rescue Guide on where to Look for Land, Air, and Water; dbS Productions: Charlottesville, VA, USA, 2008.

40. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); Honolulu, HI, USA; 2017. p. 936–44.

41. Fang P, Shi Y. Small object detection using context information fusion in Faster R-CNN. In: Proceedings of the 2018 IEEE 4th International Conference on Computer and Communications (ICCC); Chengdu, China; 2018. p. 1537–40.

42. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition; Miami, FL, USA; 2009. p. 248–55.

43. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25; Lake Tahoe, NV, USA; 2012. pp. 1097–105.

© 2025 The Author(s). | Published by Agro Intelligence Research Press

44. Ruby U, Yendapalli V. Binary cross entropy with deep learning technique for image classification. Int J Adv Trends Comput Sci Eng. 2020;9:5393–7.

© 2025 The Author(s). | Published by Agro Intelligence Research Press