International Journal of Computer Vision (2019) 127:1256–1278
https://doi.org/10.1007/s11263-019-01177-1

Deep Learning Approach in Aerial Imagery for Supporting Land Search and Rescue Missions

Dunja Božić-Štulić[1] · Željko Marušić[2] · Sven Gotovac[1]

Received: 7 May 2018 / Accepted: 9 March 2019 / Published online: 27 March 2019

Abstract

In this paper, we propose a novel approach to person detection in UAV aerial images for search and rescue tasks in Mediterranean and Sub-Mediterranean landscapes. Person detection in very high spatial resolution images involves target objects that are relatively small and often camouflaged within the environment; thus, such detection is a challenging and demanding task. The proposed method starts by reducing the search space through a visual attention algorithm that detects the salient or most prominent segments in the image. To reduce the number of non-relevant salient regions, we selected those regions most likely to contain a person using pre-trained and fine-tuned convolutional neural networks (CNNs) for detection. We established a special database called HERIDAL to train and test our model. This database was compiled for training purposes, and it contains over 68,750 image patches of wilderness acquired from an aerial perspective as well as approximately 500 labelled full-size real-world images intended for testing purposes. The proposed method achieved a detection rate of 88.9% and a precision of 34.8%, which demonstrates better effectiveness than the system currently used by Croatian Mountain search and rescue (SAR) teams (IPSAR), which is based on mean-shift segmentation. We also used the HERIDAL database to train and test a state-of-the-art region proposal network, Faster R-CNN (Ren et al. in Faster R-CNN: towards real-time object detection with region proposal networks, 2015. CoRR arXiv:1506.01497), which achieved comparable but slightly worse results than those of our proposed method.

Keywords Convolutional neural networks · RCNN · Salient object detection · Unmanned aerial vehicles (UAV) · Search and rescue · SAR image database

1 Introduction

In Croatia, Bosnia and Herzegovina (BiH) and Montenegro, responsibility for search and rescue of a missing or lost person falls under the jurisdiction of the Mountain Rescue Service. This service is specially equipped and trained for such missions. The largest number of missions are conducted on hilly, karst Mediterranean terrain, characterized by low but not dense vegetation. A search and rescue (SAR) mission begins with a call for help and ends when the missing or lost person is found-hopefully alive but possibly dead (Koester 2008). In the relevant search and rescue literature (Koester 2008), the terms 'missing' and 'lost' are not equivalent; however, in this paper, we use them interchangeably to denote the person being sought. After receiving a call, the SAR team collects all available information about the lost persons, including their gender, age, descriptions, times and locations when they were last seen, clothing and items carried and their health and psychological conditions, to mention just some of the valuable information that can help search planners to categorize the lost people into corresponding subject types. For each subject type, there are certain statistical patterns of behaviour from which it is possible to gain insight into the likely actions of the searched person. Together with factors such as type of terrain, degree of afforestation, weather conditions and time of day, the SAR leader creates a probability map of the area or

Communicated by Dr. Jason J. Corso.

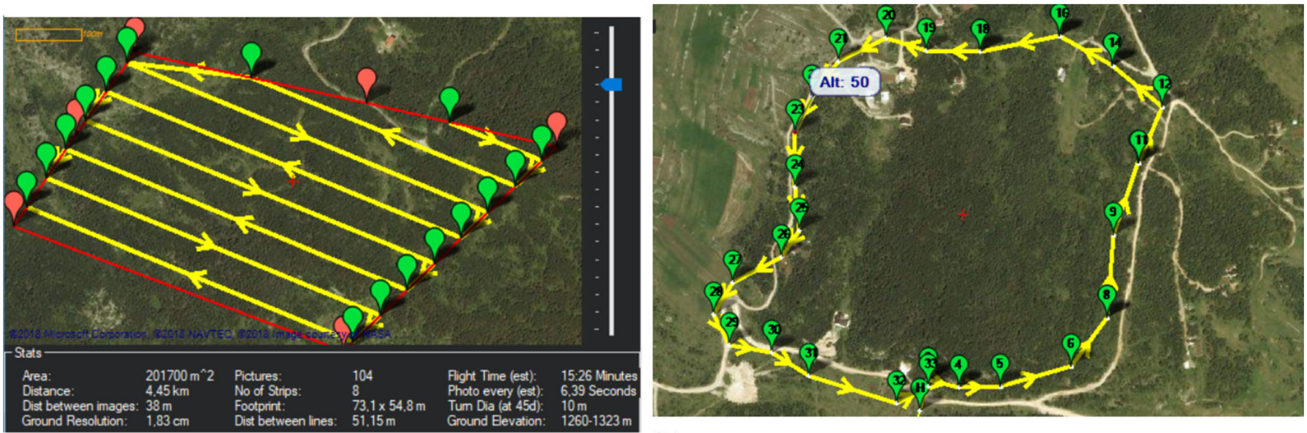
✉ Dunja Božić-Štulić
dgotovac@fesb.hr

Željko Marušić
zeljko.marusic@fpmoz.sum.ba

Sven Gotovac
gotovac@fesb.hr
https://www.fesb.unist.hr/

[1] Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, University of Split, Ruđera Boškovića 32, Split, Croatia

[2] Faculty of Science and Education, University of Mostar, Trg hrvatskih velikana 1, Mostar, Bosnia and Herzegovina


©2018 Microsoft Corporation, ©2018 NAVTEQ, ©2018 Image courtesy of NASA
Stats
Alt: 50

Area:	201700 m^2	Pictures:	104	Flight Time (est):	15:26 Minutes
Distance:	4,45 km	No of Strips:	8	Photo every (est):	6,39 Seconds
Dist between images:	38 m	Footprint:	73,1 x 54,8 m	Turn Dia (at 45d):	10 m
Ground Resolution:	1,83 cm	Dist between lines:	51,15 m	Ground Elevation:	1260-1323 m

(a) Mapping an earth section with a grid search;

(b) Following a path-patrolling and mapping road and trail sections

Fig. 1 Search scenarios for person detection from UAV imagery

'map of likelihood' where the missing person could be potentially located. The SAR team has various resources at their disposal: trackers, searching dogs, unmanned aerial vehicles and sometimes even military or police helicopters; thus, the goal is to optimally allocate available resources to each specific search task. For comparative advantages and practical reasons, UAVs have been used for years in search missions to cover large search areas rapidly and provide access to remote or difficult-to-reach locations (Yeong et al. 2015). UAVs have been especially effective for difficult-to-reach Mediterranean karst landscapes. The Croatian Mountain Rescue Service has been using UAVs on their missions for several years. According to the available SAR data, more than 50% of lost/missing hikers, hunters and children are found within a 1.8 km radius from the initial planning point (IPP) (the last known location of the missing person or the location where the person was last seen). According to SAR procedures, the first phase involves searching roads, paths and the area surrounding the IPP. At this stage, a UAV is programmed to inspect (photograph) the surrounding roads and paths or the terrain of interest. Special Mission Planner software is used to program the UAV. Figure 1 presents possible usage scenarios in SAR missions: Fig. 1a mapping an earth section with a grid search; Fig. 1b patrolling and mapping. Based on the survey grid parameters in the Mission planner, our custom-made UAV equipped with a compact 12-Mpixel camera (Canon S120) can easily map and photograph more than 20 hectares of terrain at a spatial resolution of 2 cm in less than 16 min while maintaining an altitude of 50 m. Depending on the weather conditions, the UAV records between 120 and 160 images (500–700 MB of data) per flight. It is not uncommon to use multiple UAVs. In the subsequent mission stages, the remaining terrain must be searched in accordance with the SAR mission plan.

All the acquired images are screened by an expert on the search team to detect missing persons or find other useful traces. Person detection from high spatial resolution images such as in Fig. 2, in which the target objects are relatively small and often camouflaged within the environment, is a challenging and demanding task. Based on the Croatian SAR team's experience, inspecting a single image (depending on its complexity and the display size) requires between 15 and 90 s. Images are often analysed in the field; however, when appropriate network capacity is available, they are sometimes uploaded to the cloud and processed remotely in parallel by multiple experts. Inspecting images for person detection is a slow, exhausting and tedious process that consumes excessive human resources and is subject to error. To address the above problem, a method for detecting people from aerial images based on a mean-shift algorithm (Turić et al. 2010) was developed. This method significantly facilitated the image analysis process by suggesting potential suspect locations in images; then, the search team experts could perform further visual inspections. This approach achieved a satisfactory level of detection, but it had the drawback of producing relatively significant number of false detections that were counterproductive to the visual inspection process. To overcome these disadvantages, in this study, we developed a more advanced model based on deep learning. This paper presents our new method, which has improved and enhanced the visual inspection process by utilizing saliency for region proposal and a CNN for classification. Our method functions as a support mechanism to aid in possible detection or to suggest potential locations for missing or lost persons from UAV-acquired imagery. We compare our proposed method with the existing method based on the mean-shift algorithm as well as with the state-of-the art Faster R-CNN algorithm.

The three main contributions of this paper are as follows.

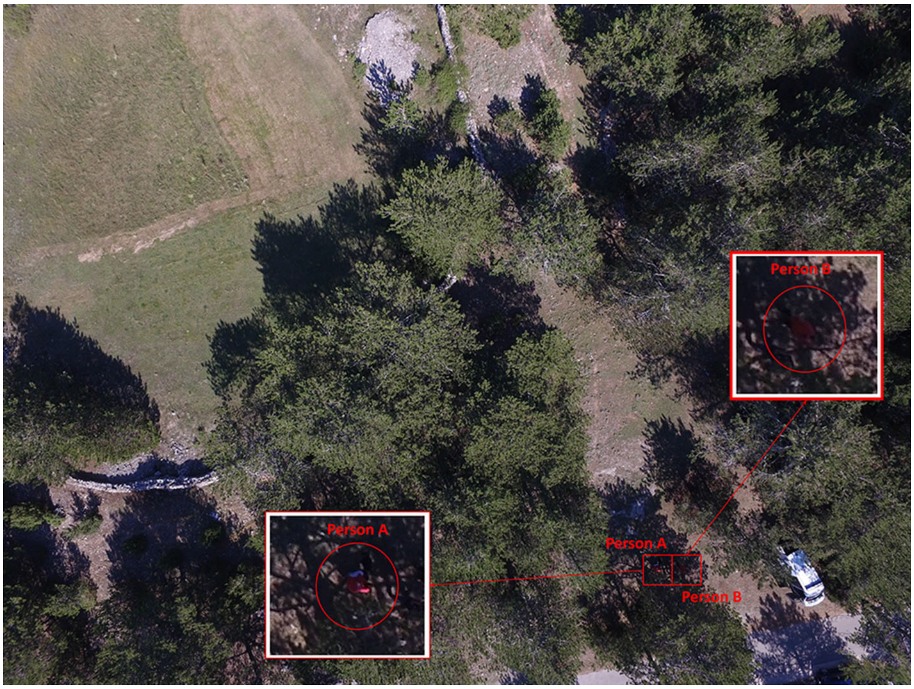


Fig. 2 Example UAV image from a SAR exercise mission

- We design a model to detect or suggest possible missing-person (small objects) locations based on saliency and deep learning models.
- We compile a public available image database of UAV imagery of lost or missing persons in natural environments called HERIDAL.[1]
- We analyse and compare the proposed model with one of our earlier solutions used by the Croatian Mountain Service on SAR missions, as well as with a state-of-art solution for region proposal and classification R-CNN.

The remainder of the paper is organized as follows. Section 2 provides a brief review of the relevant literature. Section 3 describes the proposed model and our methodology. The database of UAV imagery of lost or missing persons in natural environments intended for CNN training and testing is presented in Sect. 4. Section 5 describes the experiments and the results of the study as well as comparisons with the method based on the mean-shift algorithm and Faster R-CNN. Finally, Sects. 6 and 7 present a discussion and conclusions, respectively.

2 Related Work

Person detection from an aerial perspective is a challenging task compared to other object detection problems (for example, detecting vehicles). The wide range of person appearances resulting from changing articulated poses, clothing, lighting and background (Enzweiler and Gavrila 2009) is one challenge. However, many successful models have been created and tested that can identify objects-particularly moving objects-and these have been specifically tailored to address automotive industry issues such as pedestrian avoidance or applied to the security and surveillance domain. The simplest models involve background subtraction, which can detect moving regions across two consecutive video frames. These models are most often applied to surveillance systems, where the camera is static. However, these models are generally unsuitable when the object of interest is not moving or for object detection from still images. One early study from Viola et al. (2003) built an efficient moving-person detector that used AdBoost to select the best features from integral image representations and trained a chain of progressively more complex region rejection rules based on Haar-like wavelets and space-time differences. In recent years, based on the similarity of pedestrian detection to other generic object detection and classification tasks, deep learning models have become increasingly attractive. Some deep learning models are hybrids that combine traditional, hand-crafted features with convolutional features (Tian et al. 2015; Hosang et al. 2015), while others are pure CNN models (Zhang et al. 2016).

Person detection from UAV imagery introduces an additional spectrum of issues such as the small average size of a human body observed from an aerial platform, rapid platform motion and image instability; in combination, these issues increase the difficulty of the person detection task. Additionally, the standard approaches to pedestrian detection generally address large object sizes within an image together

[1] The data set has been published on IPSAR website, http://ipsar.fesb.unist.hr under the page “HERIDAL” or direct link: http://ipsar.fesb.unist.hr/HERIDAL%20database.html.

with texture and shape information. Several authors have proposed bimodal systems that use thermal and optical imagery from UAVs to achieve a better detection rate. Anna Gaszczak and Breckon (2011) proposed using both thermal and visible imagery for aerial reconnaissance and surveillance. Their main goal was to develop a real-time person and vehicle detection system by fusing the two image sources. For vehicle detection, they trained several independent cascaded Haar classifiers to detect vehicles in different orientations based on their lighter or darker colours in optical images. After detecting a vehicle within an optical image, they searched for confirmation in the thermal imagery. In contrast, for person detection, they first trained cascaded Haar classifiers on human thermal signatures; then, they achieved secondary detection confirmation using a multivariate Gaussian shape matching technique on regions from optical images.

Rudol and Doherty (2008) also used thermal and visible imagery to find persons lying or sitting on the ground in video sequences. They first identified high-temperature regions corresponding to human body silhouettes. The corresponding regions were then analysed in the visible spectrum using a cascade of boosted classifiers working with Haar-like features.

A significant number of papers have relied on UAVs equipped with thermal cameras as the main image acquisition source. However, in Mediterranean regions, particularly the karst areas, which are characterized by very hot summers, this type of device is not quite suitable. Most searches in the area of Dalmatia occur in the summer when the external temperatures and ground thermal radiation exceed the temperature of the human body. Search missions are generally not conducted at night because of the inaccessibility of the terrain, the danger of encountering wild animals and many other unforeseeable situations that may endanger the seekers themselves. Additionally, most state regulators forbid night flights, which is the period when thermal cameras are most effective. Consequently, in this study, we exclusively used optical cameras for image acquisition as well as in our research team's previous papers throughout the IPSAR project. A conference paper (Turić et al. 2010) from our department published by Turic, Dujmic and Papic proposed a method based primarily on the mean-shift segmentation algorithm. After segmentation tuned for small segments, the authors decided to use heuristic rules such as the sizes of segments and clusters to make decisions. The mean-shift algorithm was selected primarily because it had demonstrated good results regarding stability and segmentation quality. To reduce the high computational requirements and the quadratic computational complexity of the algorithm, they decided to modify and use two-stage mean-shift segmentation, which resulted in only a minor loss of accuracy. The pseudocode for this algorithm is provided in Appendix A. Musić et al. (2016) used the aforementioned detection model to conduct performance comparisons of the system on compressive-sensing-reconstructed images and original images, focusing primarily on image quality and information exchange. In Gotovac et al. (2016) the authors tried different approaches, applying and analysing various salient detection algorithms to detect lost persons. The conclusions from this paper are partly used in our new proposed model.

3 Proposed Model and Methodology

A number of different algorithms exist to detect objects in images. Using a sliding-window is the most popular and simplest approach and has been proven to be effective in many applications; however, this technique is not particularly efficient when used with CNN models as classifiers. The sliding window approach uses an exhaustive search to determine the locations of objects in an image, which results in a large number of image areas that must be processed to extract features for classification. CNN models are generally more computationally expensive than using generated handcrafted features; consequently the whole process can be rather slow. One solution is to use cascaded deep networks and fast features, as was done in Angelova et al. (2015), or to use a variant of region-proposal-based CNN frameworks for object detection, which was the approach taken by Girshick (2015), Girshick et al. (2013) and Ren et al. (2015). Recent advances in object detection have mostly been driven by the success of region-proposal-based or region-based R-CNNs. In their simplest forms, these models are composed of three consecutive parts. The first part reduces the feature space and selects the regions that are most likely to include the object/s of interest. Various techniques have been used to select the most promising regions; these techniques range from segmentation techniques, to saliency methods, or even to small separately trained CNN networks specially designed to identify candidate regions. After selecting the best candidates, in the second part, a computationally expensive CNN is used to extract features, and the third part applies a classifier that evaluates and performs classification. In this paper, we adopted a region-based solution that resembles the R-CNN framework pipeline (Girshick et al. 2013). An overview of the proposed model is shown in Fig. 3. To select candidate regions that possibly contain people and to obtain the relevant locations in an image, we modified the salient object detection algorithm that was described and published in Imamoglu et al. (2013). As do most algorithms in this field, this salience detection approach produces a grey-scale salience map that includes the most prominent objects. To further simplify the image, we used threshold and morphological operations to produce a binary map containing a limited number of blobs. These blobs are then further winnowed out by considering the expected size of the target object with respect to the UAV's

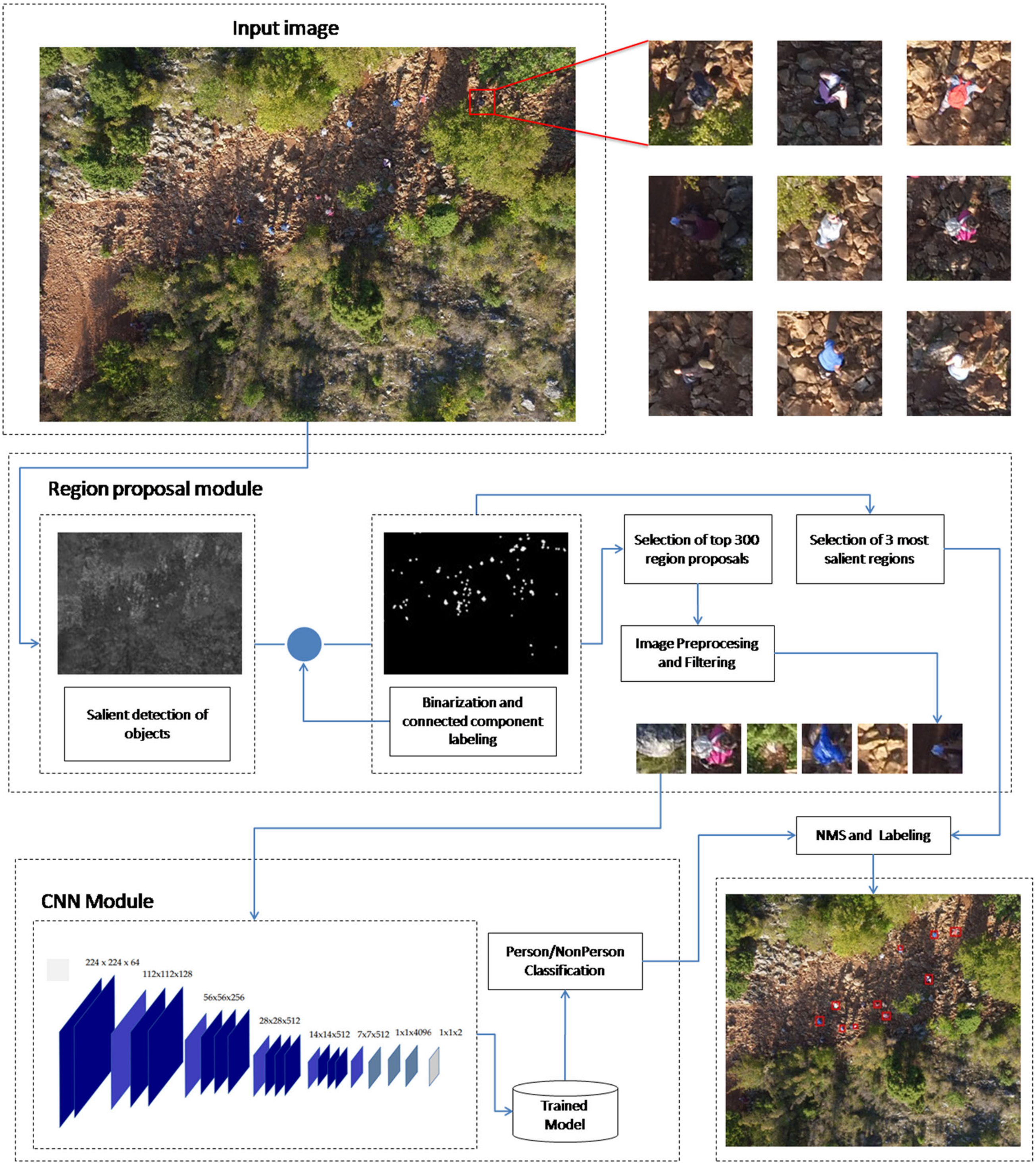

Input image
Region proposal module
Salient detection of objects
Binarization and connected component labeling
Selection of top 300 region proposals
Selection of 3 most salient regions
Image Preprocesing and Filtering
NMS and Labeling
CNN Module
224 x 224 x 64
112x112x128
56x56x256
28x28x512
14x14x512
7x7x512
1x1x4096
1x1x2
Person/NonPerson Classification
Trained Model

Fig. 3 Architecture of proposed model

flight altitude and camera characteristics. The binary map serves as a mask that allows extraction of the image parts used as input to the CNN classification module. The network processes the candidate patches using several convolutional (conv) and max pooling layers to produce several convolution feature maps that are subsequently used to produce fixed-length feature vectors classified as 'person' or 'non-person'. The final detection map shown in Fig. 3 includes some of the most prominent regions (3 per image) and was produced by skipping the classification model's evaluation. More detailed elaborations of the proposed model can be found in the following subsections.

3.1 Region Proposals Using Saliency Object Detection

One of the most important features of the human visual system (HVS) is that it derives important and compact information from natural scenes using a process called visual attention. Because the surrounding environment includes an excessive amount of information, the visual attention mechanism reduces the redundant data, concentrating only on the information that benefits perception during the selective attention process (Treisman and Gelade 1980; Koch and Ullman 1987; Imamoglu et al. 2013). In this way, humans can quickly locate the most important parts of a scene - the parts that stand out relative to their neighbours and thus capture our attention.

These prominent, dominant or interesting objects are also called ‘salient objects’. Computational models that imitate biologically inspired processes and detect objects of interest are called salience detection models. Many studies have attempted to build computational models to simulate this mechanism (Borji et al. 2014); however, most studies have examined and tested on high-quality and high-resolution colour images with minimal noise in which where the salient object is the main subject or occupies the a substantial part of the image (Sokalski et al. 2010). In contrast, our subjects of interest occupy an extremely small area on of the image, and; therefore, some the classical (non-deep learning) state-of-the-art models typically give yield only the average or not unsatisfactory results. However, Leroy et al. (2014) compared the results of nine salient detection algorithms on the Jian Li database (Li et al. 2016) for objects with three size categories (large, medium and small). Based on this paper, an algorithm based on wavelete transformation (WT) (Imamoglu et al. 2013) demonstrated excellent results in the detection of small objects. That conclusion was confirmed in one of our earlier papers (Gotovac et al. 2016), where we compared some bottom-up salient object detection algorithms on the images obtained from the UAV aircraft. Because this algorithm is very highly suitable for detecting smaller salient parts portions of the an image, we included WT (Imamoglu et al. 2013) as the base of our region proposal part/module to simplify the image and find regions of interest or candidates. An overview of WT model is shown in Fig. 4.

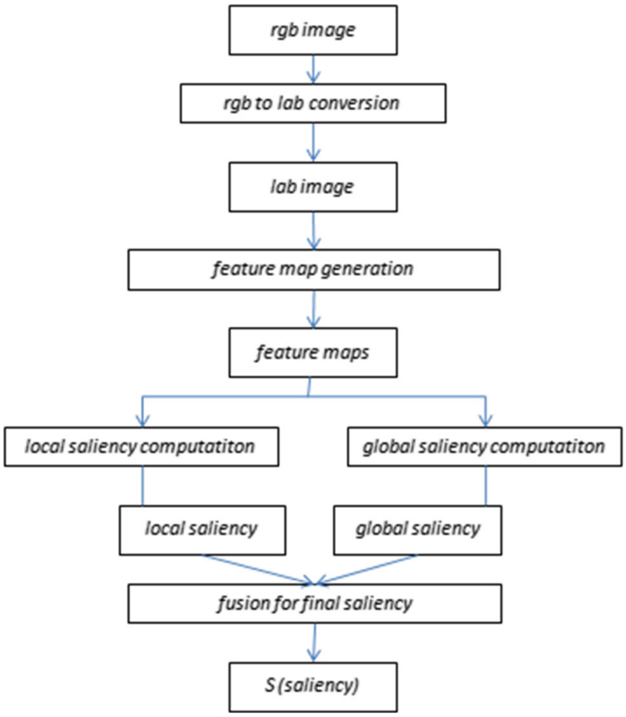


Fig. 4 Framework of saliency detection model based on wavelet transform

In the first phase, RGB images are converted into CIE Lab colour space due to the fact that because Lab that colour space is more uniform and better approximates similar to human perception. Then, a Gaussian low-pass filter is used to remove the noise from input colour image I:

$$I_n = I * G_{mxm} \tag{1}$$

$\mathrm{G}_{\mathrm{mxm}}$ represents a 2D Gaussian filter in which we set m = 5. The filtered image I_{n} channels are then normalized to the range [0, 255]. Wavelet decomposition is used to extract oriented details (horizontal, vertical and diagonal) from the multi-scale perspective on in the normalized channels. Wavelet decomposition enables high spatial resolution with higher frequency components and low spatial resolution with lower frequency components without losing detail information loss in detail during the decomposition process. The sub-bands of the image formed by WT for a number of levels using Daubechies wavelets (Daubechies 1992) $1, \ldots, N$ wavelets (db5) as follows:

$$[A_N^c, H_s^c, V_s^c, D_s^c] = WT_n(I_n) \tag{2}$$

where N represents the maximum scaling number for the WT decomposition process and the level with the coarsest resolution. S represents the resolution index s $\in \{1, \ldots, N\}$, and c represents the image colour channels c $\in \{L, a, b\}$. H_s^c, V_s^c, D_s^c are the wavelet coefficients of the horizontal, vertical and diagonal details for the given s and c, respectively, and A_N^c is the an approximation component at the coarsest level, which is not used during feature extraction. Feature maps $f_s^c(x, y)$ are created during the inverse wavelet transformation (IWT) of the wavelet coefficients for the sth level decomposition for each image sub-band c (the approximation component A_N^c is used during feature extraction) as follows:

$$f_s^c(x, y) = \frac{(IWT_s(H_s^c, V_s^c, D_s^c)^2}{\eta} \tag{3}$$

where η scaling factor to limit the range of feature values and is necessary to avoid huge variations in the covariance matrix among the feature maps during the computation of the global saliency map. After calculating the feature maps, a local saliency map $s_l(x, y)$ is created by fusing the feature maps at each level linearly using the formula given below:

$$s_l(x, y) = \left(\sum_{s=1}^{N} argmax(f_s^L(x, y), f_s^a(x, y), f_s^b(x, y)) \right) *G_{mxm} \tag{4}$$

A global saliency map is calculated on from a 3 × N feature vector (3×3 channels and N-level wavelet-based features for each channel) from all the feature maps using a probability density function (PDF), that calculates the likelihood of the features appearing at a given location. The PDF in multi-dimensional space is given by

$$p(f(x, y)) = \frac{1}{(2\pi)^{n/2} |\sum|^{1/2}} * xe^{-1/2(f(x,y)-\mu)^T \sum^{-1}(f(x,y)-\mu)}$$

with

$$\sum = E[(f(x, y) - \mu)(f(x, y) - \mu)^T] \tag{5}$$

where μ is a mean vector containing the mean of each feature map, T is the transpose operation, $\sum$ in (5) is the covariance matrix; $|\sum|$ is the determinant of the covariance matrix. Using the PDF from (5), the global saliency $s_g(x, y)$ map can be computed as follows:

$$s_g(x, y) = (log(p(f(x, y))^{-1}))^{1/2} * G_{mxm} \tag{6}$$

In the final step, the local and global salient maps are combined using a modulation function M as a non-linear normalization function ($M(.) = \frac{(.)^{ln\sqrt{2}}}{\sqrt{2}}$)to produce final salience map, $S(x, y)$

$$S(x, y) = M(S_l(x, y)xe^{s_g(x,y)})xG_{mxm} \tag{7}$$

As shown in (4) (6) and (7), each resulting salient map is filtered with a 2-D Gaussian low-pass filter to obtain a set of smooth maps. In our work, we made several changes and adjustments to this algorithm to enhance its performance and improve its execution speed. In the original paper, the authors enhanced the final saliency map based on Gestalt law principles. These principles can be interpreted in such a way that locations around the focus of attention (FOA) gain more attention than those that are further away. Thus, the more salient points in the saliency map are enhanced, while the points more distant from the salient point are suppressed. However, we did not use these principles in our evaluations because it showed no additional benefits on our set of images; therefore, we disregard this technique to enhance the performance. One reason we chose this algorithm for region proposal instead of one of the pure segmentation methods, such as the mean-shift algorithm, was its linear asymptotic execution time. However, in its original form, this algorithm applies an extensive number of calculations to a larger amount of data, which makes it impractical for larger-sized images. Consequently, it was necessary to make certain considerations and adjustments. The algorithm's main drawbacks are its memory requirements and the amount of data needed to calculate the global saliency map or probability density map. After obtaining the feature maps from (3), we need to concatenate them for every channel and at every level of decomposition. For example, calculation of a global saliency map for an RGB image with a size of 4000 × 3000 requires a 12-level wavelet decomposition that generates 12 feature maps per channel (36 feature maps overall). Each feature map has the same width and height as the original image; therefore, we would need to calculate 36 * 4000 * 3000 or 432 million features. To store this much information in memory, we would need approximately 1728 MB of single-precision floating point memory space. However, using every feature map and every decomposition level to calculate the probability density map is overkill. Experimentally, we found that no accuracy was lost by using only every third feature map per channel; this approach results in 12 * 4000 * 3000 or 144 million features, which significantly reduces the performance impact. As a proof of concept, the first step was to implement the algorithm in MATLAB. Table 1 shows the execution time with respect to the size of the image and the number of features generated by the modified version of the algorithm. The Data/Time Ratio row shows that the algorithm behaves linearly with respect to data growth. Although it may seem from Table 1 that the algorithm is uncompetitive regarding execution speed, we note that the original MATLAB implementation was not optimized for multi-core processors or GPUs. By implementing the algorithm in C++ and CUDA, we managed to achieve a 38x speedup on a computer with an Intel Xeon (6 Core) processor and an Nvidia Titan GTX graphic card, which resulted in an execution time of 1.89 s for an RGB 4000 × 3000-pixel image.

3.2 Deep Learning and Feature Extraction Using a CNN

Deep convolution neural networks (DCNNs) have recently shown outstanding performances on image classification tasks. In addition to image classification, DCNNs have also

Table 1 Time analysis of the modified WT algorithm implemented in MATLAB and implemented and optimized in C++

Image size	500 × 375	1000 × 375	1000 × 750	2000 × 750	2000 × 1500	4000 × 1500	4000 × 3000
Numb. of features (mil)	1.69	3.38	6.75	13.50	27.00	72.00	144.00
Execution time(s)—Matlab	0.93	1.81	3.54	7.01	13.87	35.18	71.45
Data/time ratio—Matlab	1.82	1.87	1.91	1.93	1.95	2.05	2.02
Execution time(s)—CUDA	0.023	0.046	0.091	0.179	0.36	0.93	1.89
Data/time ratio—CUDA	73.55	73.91	74.12	75.31	75.82	77.25	76.19

been used in object detection tasks to provide precise locations of the detected objects.

DCNNs are composed of several layers, each containing linear or non-linear operators that are jointly trained in an end-to-end fashion to solve specific tasks. Certain deep architectures have contributed significantly to the field of deep learning and some architectures have become standards [e.g., LeNet5 (Lecun et al. 1998), AlexNet (Krizhevsky et al. 2012), VGG-16 (Simonyan and Zisserman 2014), GoogLeNet (Szegedy et al. 2015) and ResNet (He et al. 2015)]. In our paper, we chose the visual geometry group (VGG) convolutional neural network proposed by Simonyan and Zisserman (2014) at the University of Oxford. They have suggested various models and configurations of deep convolution neural networks and presented one of their proposals to ILSVRC-2013. This model, also known as VGG-16 (because it contains 16 layers), became popular for achieving TOP-5 accuracy of as much as 92.6% on the ILSVRC Image classification task and showed that network depth is a critical component in achieving good performance.

The main difference between the VGG-16 model and its predecessors (LeNet and AlexNet) is that VGG-16 uses many convolutional layers with small receptive fields in the first layers of the network. It features an extremely homogeneous architecture that performs only 3 × 3 convolutions and 2 × 2 pooling from the beginning to the end. This approach reduced the parameters and increased the network's nonlinearity, resulting in a model that is easier to train.

CNN models with various architectures are primarily used to learn features from data that constitute a main-but not the only-building block in successful generic object detection frameworks. As previously mentioned, we used a pipelined approach similar to that of R-CNN (Girshick et al. 2013) which significantly improved the detection performance in terms of mean average precision (mAP) compared to models that did not use CNNs. This system consists of three parts. The first part generates region proposals using a selective search. The second part uses the CNN to extracts features from every proposed region. The third part uses an SVM to perform classification. The Fast-RCNN model was presented in Girshick (2015) and it improved the training and testing speed of the original R-CNN and its detection accuracy by sharing the CNN computation among all the region proposals. In this model, an image is first input into a CNN to create a convolution feature map; then, the Region of Interest (RoI) pooling layer extracts a feature vector for each region proposal. The feature vectors are fed into fully connected layers. Finally, the model produces soft-max class-probability estimates and bounding boxes for each detected object. The bottleneck of these two proposed systems lies in the first part or region proposal stage, where a selective search (SS) greedily merges super-pixels based on engineered low-level features, which is relatively time-consuming operation. In Ren et al. (2015) the authors addressed this issue by using a Region Proposal Network (RPN)—a separate neural network responsible only for suggesting candidate regions. In the RPN module, a small network slides over the convolution feature map with multiple anchors at each sliding window location and then outputs bounding boxes (region proposals) that are input to the Fast R-CNN detector for inspection. The proposed model is known as Faster R-CNN and considerably reduced the computational requirements of the overall inference process.

In general, we can say that all three of the above models efficiently classify object proposals using deep convolutional networks and have achieved state-of-the-art status on standard classification and detection data-sets, they have become the de facto standards of this field. However, the standard image databases mostly contain objects that are prominent, large, and occupy a sizable proportion of the total image area. Therefore, in their original form these frameworks did not yield similar performance with regard to the detection of relatively small objects. The authors of Chen et al. (2017) evaluated this issue and empirically determined that the selective search and edge box approaches work well in generating proposals for large objects in the PASCAL VOC dataset, but were unsatisfactory for generating small object proposals. Even with exhaustive search and 2000 object proposals per image, the recall rate was lower than 60%.

Region proposals by an RPN present similar behaviours and do not provide results with the same accuracy, as reported in Yuan et al. (2017); Chen et al. (2017) and Eggert et al. (2017). This result occurs because RPN uses hard-coded anchors with fixed scales and aspects when identifying potential regions. Therefore, to use these generic frameworks effectively for small-object detection, the problem characteristic considered in this paper, in most cases it is necessary to modify the part that generates region proposals. In this study, we decided to select regions of interest from saliency maps. These image proposals are used as input to pretrained and fine-tuned VGG-16 CNN networks using transfer learning from our dataset. The final classifications are performed using the SoftMax classifier.

4 Design and Implementation of the HERIDAL Image Database

One important consideration in object detection and other computer vision tasks is acquiring a relevant image database. Image database has been design according to the instructions presented in paper (Zendel et al. 2017). In general, image data are collected from a large variety of locations and sources, from social media activities to surveillance data. However, due to the specific problem of detecting people in SAR missions in non-urban areas, these image sources are inadequate. Additionally, it is worth mentioning that the experimental images in related works mostly consist of a small number of images or are created in simulated conditions that do not correspond to real-world scenarios. We considered it necessary to take certain theoretical and practical knowledge from the SAR literature into account when designing and collecting an image dataset that can be used for the experimental phase.

It is particularly important that the set contains as many realistic scenes as possible, including mountains, wilderness and remote terrain and that it covers most real-world situations that could be encountered in practice, such as person poses, colour of clothing, position in the environment, illumination, etc. All the data required to build such a database can be gleaned from statistics of the Croatian SAR team as well as from the literature, especially from the book (Koester 2008) that represents the starting point of SAR theory. In that book, the author processed and interpreted data obtained from the International Search and Rescue Incident Database (ISRID), which contains 50,692 SAR cases. According to Syrotuck and Syrotuck (2000) and Koester (2008), subject types can be classified by age, mental status and activity into a dozen broad categories. Some of these categories involve children (various age groups), hunters, mushroom gatherers, hikers, climbers, irrational, mentally challenged and despondent people. Each category is characterized by specific behaviour patterns that can manifest when a person is lost and that can be helpful in predicting the missing person's potential location and pose in the environment. For example, an irrational person tends to move in a straight line until they encounter obstacles such as bushes, trees or reach water. Depressed people as a subject category include people who have been depressed and those with a high possibility of suicide or who have expressed the intent to commit suicide. Most just seek to get out of sight so they cannot be considered lost persons; sometimes, they intentionally hide themselves, do not answer calls or actively avoid the search team. Climbers and mountaineers often travel considerable distances to ascend prominent peaks or to climb rocks. The biggest risks for these two categories are over- or underestimation of the terrain difficulty or the time required to complete the climb. The second biggest risk is becoming stranded and bad weather. Trauma and injuries from falls or falling debris are also frequent. Hunters and berry and mushroom pickers generally get lost in the wilderness while travelling cross-country. Preoccupied with their activities, they often fail to pay attention to changing terrain, weather conditions or the passage of time. Children frequently become lost because they take shortcuts that may in actuality represent a longer route. They are also commonly involved in fantasy play, exploring or adventuring, and in most cases, they are drawn to wilderness areas. When lost, they use a trail/road following strategy. Additionally, children may be hiding intentionally to avoid punishment, to gain attention or simply sulking. From these examples and the statistical data presented in Table 2, we can draw some conclusions about potential missing-person locations.

As shown in Table 2 for the selected types of subjects, most people were found along roads, drains or other linear structures (e.g., trails, railroads, or pipelines), followed by structures (enclosed structures or shelters), and finally, in bushes/forests. However, the table does not include all the types of subjects, not does it apply to all climatic zones (only moderate climate zones). Such information as well as information concerning the average distance from the IPP for a particular subject type are used by search planners primarily to determine the probability of area (POA) and thus, position a search team in the most probable location as soon as possible. Information about locations where people can potentially be found was highly important in designing a proper and valid image database for use in the experimental phase in this paper. The dataset needs to include open fields, roads, drainage areas, bushes and forest as well as enclosed structures or shelters.

Additionally, it would be beneficial to know the body position of injured or missing persons at the moment they were found, but these data are not available from existing statistical databases related to SAR missions. This situation is understandable because such information is not particularly important or relevant for classic SAR missions. In contrast, such information could contribute significantly to a better performance for computer vision tasks. We mentioned that depressed people may actively attempt to avoid the search team or crouch down to hide in bushes. Mountaineers or hikers can experience an accident and severe trauma that causes them to remain immobile in various lying or sitting positions. Children as a subject type have a much smaller projection and

Table 2 Finding locations for some selected subject types

Subject type	Structure	Road, Drainage, Linear	Brush, scrub, woods	Field	Rock	Other
Demented	20	36	23	14	0	7
Despondent	26	26	26	6	2	14
Climbers	0	27	9	9	27	28
Gatherers	0	80	10	0	10	0
Hikers	13	50	9	14	4	10
Hunters	8	55	14	0	2	21
Children (7–9)	29	38	15	6	1	11
Average	16	44	15	7	6	13

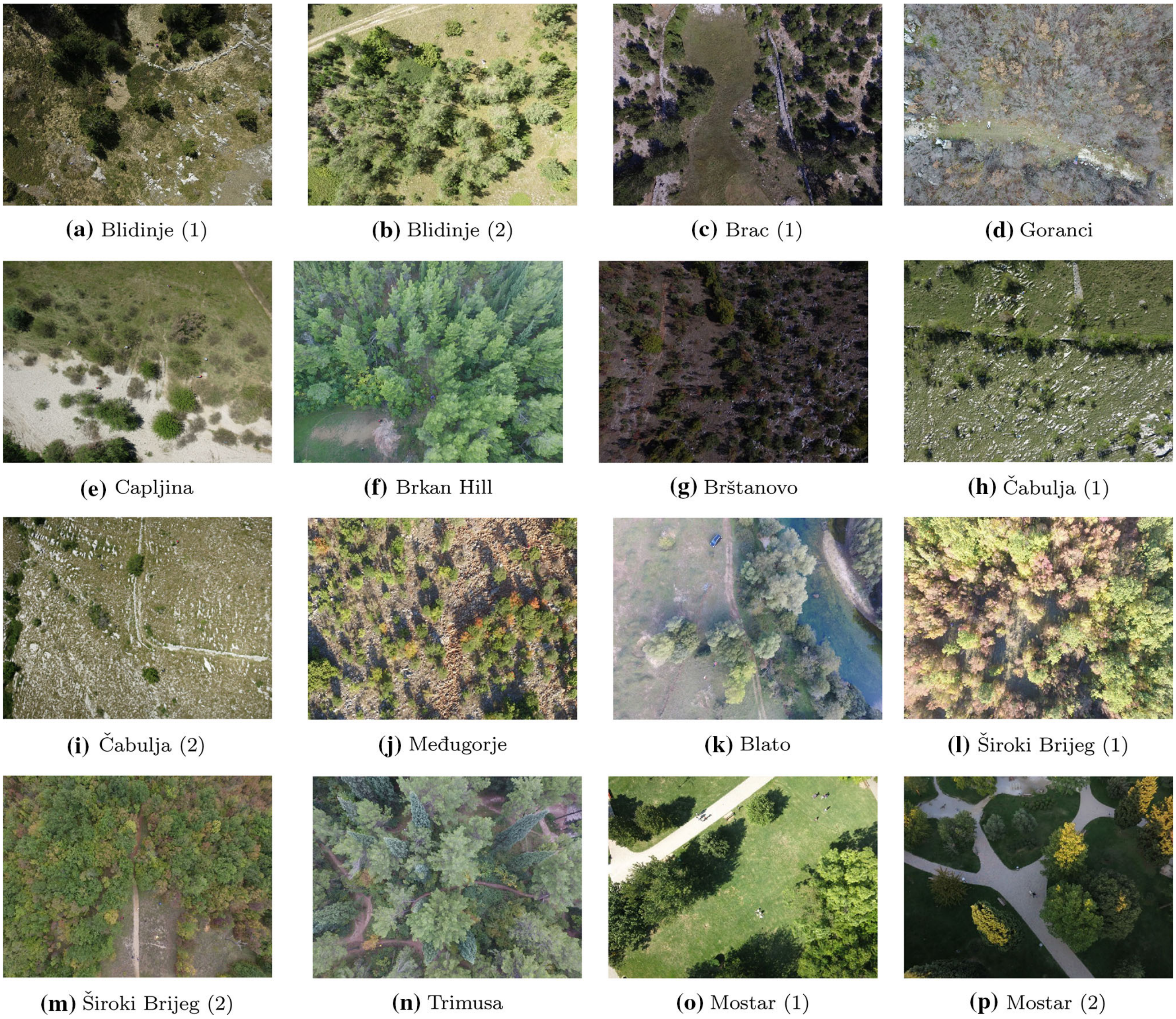
footprint in images compared to adults; they are sometimes difficult to distinguish from natural artefacts, especially in low lighting conditions. The silhouette and shape of a person stuck in the bush and trees can be severely occluded such that only a small part of the person is visible from the air, which further complicates the entire detection model.

It should also be noted that not all searches are related to only one missing person. A subject can be lost alone or in a group. The majority, approximately two-thirds (67%), of all incidents involve solo subjects. Groups of two account for 19%, and the rest are groups of 3 or more subjects. According to these data, we obviously cannot expect that every search will be organized for a solo subject. Thus, the possibility of a search for multiple subjects cannot be neglected; we need to ensure that the dataset includes images with multiple people as well as images with traces of human presence.

Figure 5 shows the variety of terrain and an excerpt of the image database on which the experimental analysis was performed. A test set of 101 images was gathered and selected from 12 locations with various terrain configurations from different locations in Croatia and BiH. The images were acquired during mountain tours, nature trips or during mountain rescue service exercises. At several locations, we used staged volunteers and students to act as missing persons. Images (vertical or nadir) were acquired at altitudes of 40 m to 65 m from a custom-made UAV equipped with a Canon Powershot S110 camera and Mavic Pro from a Phantom 3 Advanced model. Both models are equipped with sensors capable of photographing 12-Mpx images, resulting in ground resolution of 2 cm. At one location (Medjugorje), images were taken with a DJI Inspire UAV (wide-format 4000 × 2250 px). Each image contains at least one "missing" person, and some of the images contain more than 10 people. All the people in the images are adequately labelled. Table 3 lists information about the image locations, the number of images per location, the number of persons per location, as well as various person poses in the scene presented according to statistic proposed by Croatian Mountain SAR Team. Table 4 presents a distribution of the images according to lost person location given in Table 2. It should be emphasised that the person location is not easy uniquely to determine. For example, when according to statistics a person was found on the road in many cases it is not directly on the road but in nearby bushes, shrub or rocky area. Also, locating persons on the road is not a challenging task, while locating them in bushes, shrub or rocky area is incomparably more difficult.

We emphasize that the level of scene lighting is an important factor in acquiring good and usable images, and it is highly dependent on the time of day a photograph was acquired (dawn, morning, noon, sunset) as well as the meteorological conditions (sunny, cloudy, partly cloudy). These dynamic characteristics can be partly compensated for by the settings of the image capture devices (i.e., ISO, exposure, and lens aperture). However, in practice, due to the complexity of the scenes, the resulting images are overexposed or underexposed, which can result in reduced detection accuracy or an increased number of false alarms.

For example, although Fig. 5a, b show similar areas, Fig. 5b is somewhat overexposed compared to Fig. 5a. The same conclusion applies to Fig. 5i, m. However, in the first case, we are dealing with different camera settings, while in the second case, changed weather conditions resulted in a greatly altered scene. The pictures shown in Fig. 5g, p are somewhat underexposed. Figure 5h, i show micro-localities of the Cabulja mountains in spring and summer: we can clearly see changes in the vegetation (green and yellow grass). Generally, each image includes a large amount of data and details that must be inspected. In this wealth of information, natural artefacts and objects are predominant, whereas artificial or manufactured objects or missing persons are generally extremely rare. From a dataset perspective, we face a highly unbalanced set that is not suitable for standard machine learning detection models and leads us to the field of artificial intelligence/machine learning that deals with outliers, which in most cases yields poorer results. To overcome these issues, we invested a great deal of time and effort in creating a training database that includes many positive examples of persons on various terrains to



(a) Blidinje (1) **(b)** Blidinje (2) **(c)** Brac (1) **(d)** Goranci

(e) Capljina **(f)** Brkan Hill **(g)** Brštanovo **(h)** Čabulja (1)

(i) Čabulja (2) **(j)** Međugorje **(k)** Blato **(l)** Široki Brijeg (1)

(m) Široki Brijeg (2) **(n)** Trimusa **(o)** Mostar (1) **(p)** Mostar (2)

Fig. 5 Example of images from various terrains and locations: **a** Blidinje (1), **b** Blidinje(2), **c** Brač, **d** Goranci, **e** Čapljina, **f** Brkan Hill, Mostar, **g** Brštanovo, Split, **h** Čabulja (1), **i** Čabulja (2), **j** Međugorje, **k** Blato, **l** Široki Brijeg (1), **m** Široki Brijeg (2), **n** Trimusa, Mostar (1), **o** Zrinjevac (2), **p** Zrinjevac (3)

allow us to train machine learning algorithms in general and deep-learning approaches in particular. Logistically and organizationally, it would be very difficult and expensive to gather large numbers of people to simulate injured people in wilderness areas. Therefore, we gathered images of people on several occasions, as illustrated in Fig. 6. Figure 6a shows Catholic believers climbing on Apparition Hill in Medjugorje (12/08/2017); Fig. 6b shows mountaineers climbing at Mount Cvrsnica; Fig. 6c shows a mass meeting at Kupres (30/07/2017); and Fig. 6d shows people gathering in a park in Mostar (all locations are in Bosnia and Herzegovina). These vertical or nadir image acquisitions were acquired from a UAV equipped with a Canon Powershot S110 camera, which is characterized by a CMOS sensor capable of taking 12-Mpx images at altitudes from 30 to 40 m. These images include a large number of people (children, women, youths and the elderly) wearing a variety of clothing, in various poses and positions within the environment, and under differing lighting conditions. By cropping people from the images, we created a database containing 12,378 positive patches. However, these images did not include enough people sitting or lying down, which is a common situation in SAR missions. To address this problem, we decided to synthetically generate such positive patches. We gathered students and photographed them in various positions (standing, walking, sitting, squatting, lying down) in a local stadium. This approach placed the people on a uniform grassy background that can be easily removed from patches. Subsequently, we

Table 3 Details about image database and distribution of various person poses in the scene

Location	Abbr.	Platform	Number of images	Number of person	Standing	Sitting	Laying	Squatting
Blidinje	BLI	Custom (Cannon S120)	9	47	10	12	18	7
Brač	BRA	Phantom 3 Pro	12	22	19	3	0	0
Brkan hill	BRK	Cannon S120	12	21	5	10	6	0
Brstanovo	BRS	Phantom 3 Pro	11	21	18	0	3	0
Čabulja	CAB	Custom (Cannon S120)	8	34	4	3	16	11
Čapljina	CAP	Custom (Cannon S120)	6	19	1	1	11	6
Medjugorje	MED	Custom (Cannon S120)	6	53	49	1	0	3
Blato	MOB	Phantom 3 Pro	5	13	7	0	5	1
Široki Brijeg	SB	Phantom 3 Pro	10	24	11	3	8	2
Trimusa	TRS	Phantom 3 Pro	6	18	11	0	0	7
Zrinjevac	ZRI	Custom (Cannon S120)	6	49	29	15	3	2
Goranci	GRO	Mavic Pro	10	20	13	1	3	3
Overall			101	341	177	49	73	42

Table 4 Distribution of the images according to lost person location

Location	Abbr.	Number of images	Road and near road	Brush, scrub	Field	Rocks
Blidinje	BLI	9	6	2	1	0
Brač	BRA	12	4	8	0	0
Brkan hill	BRK	12	3	9	0	0
Brstanovo	BRS	11	11	0	0	0
Čabulja	CAB	8	0	0	0	8
Čapljina	CAP	6	4	0	2	0
Medjugorje	MED	6	0	0	0	6
Blato	MOB	5	0	1	4	0
Široki Brijeg	SB	10	5	5	0	0
Trimusa	TRS	6	6	0	0	0
Zrinjevac	ZRI	6	6	0	0	0
Goranci	GRO	10	0	10	0	0
Overall		101	45	35	7	14

inserted the person images into real environment images, creating an additional 1000 positive patches. This procedure is presented in Fig. 7.

For training purposes, we also needed negative patches. We selected negative patches using the proposed saliency detection algorithm from real images with a 4000 × 3000-pixel resolution. By applying the algorithm to images that do not contain people and cropping the most prominent salient regions, we created 19,850 negative patches. To create additional positive sample, we performed data augmentation using rotation, shear and horizontal flip operations. For the negative samples, we used only rotation by 90 degrees as a data augmentation technique. In this way, we extracted 29,050 positive samples and 39,700 negative samples. After finding the centroid/anchor of a particular blob, we created and labelled a rectangular 80 × 80-pixel bounding box around it. Examples of positive and negative patches are shown in Fig. 8.

Using the techniques described above, we established a special database called HERIDAL for training and testing purposes. HERIDAL is specifically designed to cover most of the real-world situations encountered in practice, including person poses, clothing colours, people's positions in the environment, varying illumination levels, etc. This database contains over 68,750 image patches of people in wilderness locations viewed from an aerial perspective. Of these, approximately 3000 image patches are synthetically generated; the others are cropped from real images. Additionally, the HERIDAL database contains approximately 500 labelled, full-size 4000 × 3000 pixel real-world images for testing purposes.

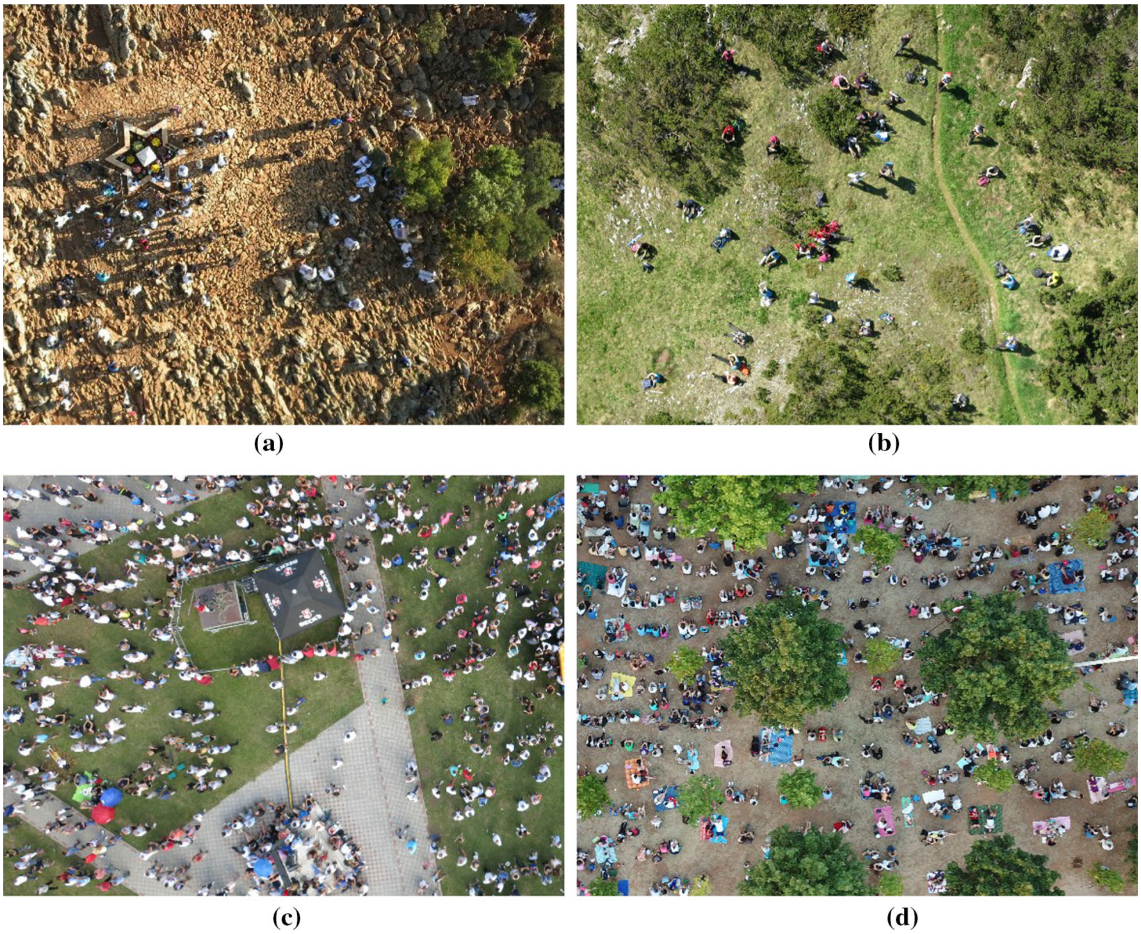


Fig. 6 Gathering patches: **a** Apparition Hill in Medjugorje; **b** Mountaineers at Cvrsnica; **c** Mass gathering in Kupres; **d** People gathered in a park

5 Experiments and Results

In this chapter, we explain the entire proposed model in detail and present the results of the experiments. Testing was performed on the aerial imagery dataset HERIDAL, which is described in Sect. 4. From a test set of 101 images, the most representative aerial images were taken at non-urban locations in BiH and Croatia. The altitude of the UAV of the platform was 45–60 m above the ground. The resolution of most of the acquired images is 4000 × 3000 pixels. At least one person is present in each image; on average, images contain 3.37 persons. We adopted three well-known measurements to properly evaluate the separate stages and the overall detection model: Recall, Precision and Accuracy.

Recall is the number of true positives relative to the sum of the true positives and the false negatives. Recall represents the percentage of people correctly detected among all the candidate regions that should have been detected as people. Here, true positive (TP) is the number of correctly detected persons, and false negatives (FN) is the number of misdetected persons (detection failures). Recall also represents the detection rate. Precision represents the percentage of correctly detected people divided by the total detected people. It includes false positives (FP), detected objects that are false alarms or are incorrectly detected as people. Accuracy is a measure of the performance of the system with regard to both correctly detecting and correctly rejecting targets. It is the sum of the true positive and true negatives relative to the total number of region proposals. True negatives (TN) represents the number of objects correctly identified as background or non-person. In this paper, we did not adopt bounding box regression to measure the precise localization of an object

5.1 Region Proposal Module (RPM) Evaluation

The region proposal module based on the salient detection algorithm when applied to a colour image produces a greyscale salient map. To suggest region proposals from the saliency maps, we first need to produce binary maps or binary masks. The key step in calculating a binary mask is binarization of the saliency map S-and the simplest way to do that is to pick a threshold that varies between 0 and 255.

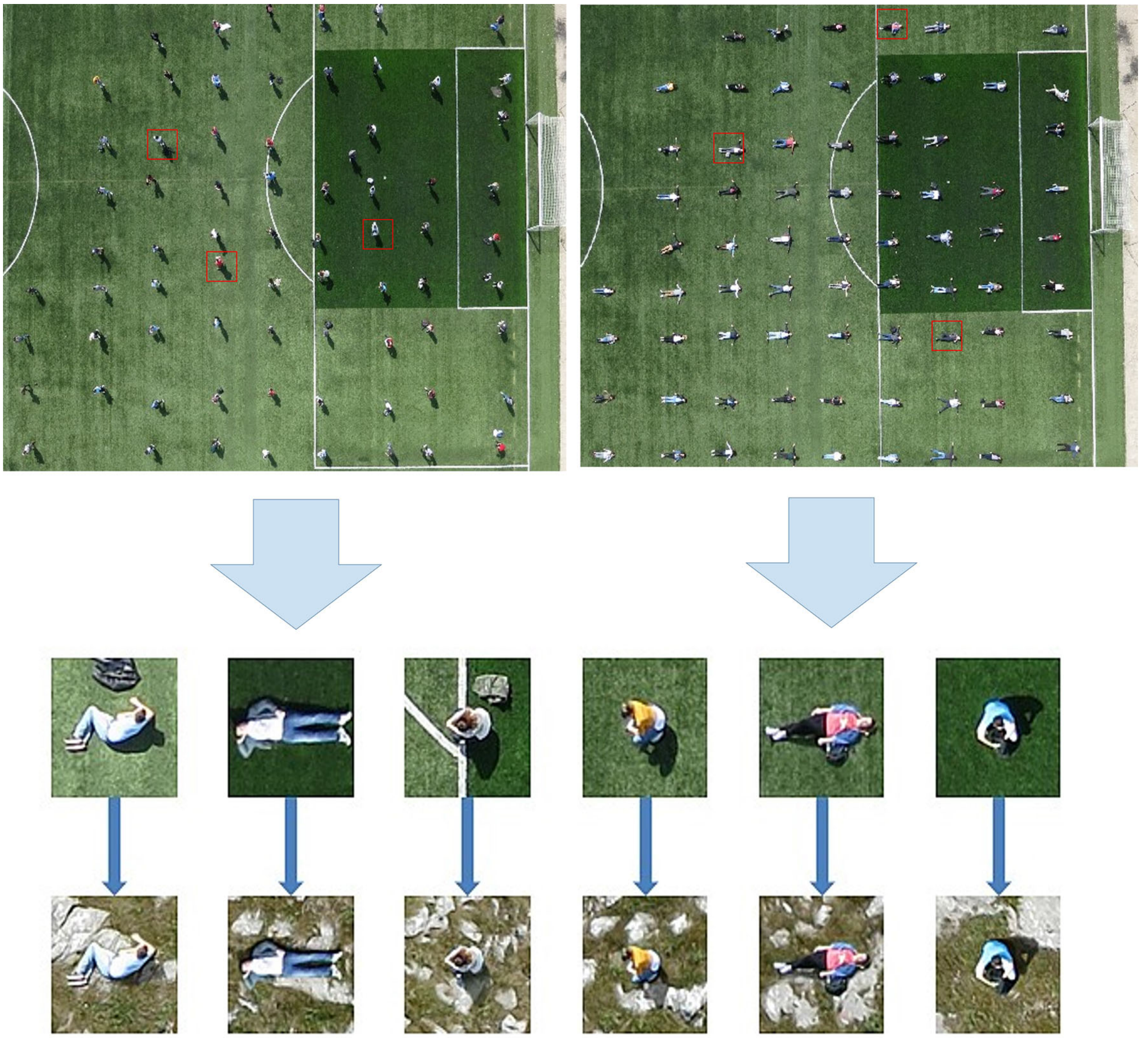


Fig. 7 Procedure of stitching human patches on various backgrounds

Then, for each threshold value, a pair of (Precision, Recall) scores are computed and used to plot a precision-recall (PR) curve. In this way, we can select the best ratio between recall and precision. Additionally, one popular choice is to use an image-dependent adaptive threshold for binarizing S, which is computed as twice the mean saliency of S:

$$T_a = \frac{2}{W x H} \sum_{x=1}^{W} \sum_{y=1}^{H} S(x, y) \tag{8}$$

Unfortunately, neither of these popular approaches are suitable for our task. Using the first approach requires manually adjusting the threshold step for each image, which is highly impractical. The second approach has not proven reliable with complex images such as ours (Fig. 12). However, this situation is quite reasonable because our task does not represent a standard saliency object detection problem. The typical saliency goal is to find one or a small number of prominent areas that occupy relatively substantial areas of the image. In contrast, our goal is to simplify a complex scene by filtering to retain only those parts that are sufficiently prominent and conspicuous. In this phase, it is crucial to focus on better recall: in other words, if we miss an important part of the image that represents a person at this early stage, the person will be irretrievably lost at the later classification stage. After performing some analysis, we found that in most cases, 300 proposals or binary blobs is the optimal number. To accomplish this task, we took an iterative approach in which we binarized the saliency map using a starting threshold value of 115 (the 45% intensity level for grey-scale images) in the first iteration. Then, the connected-component labelling operation was performed as a convenient way to determine how many blobs were present in the binary map. When the total number of blobs after labelling exceeds 300, we repeat the entire process using a higher threshold value in the binarization process. In every

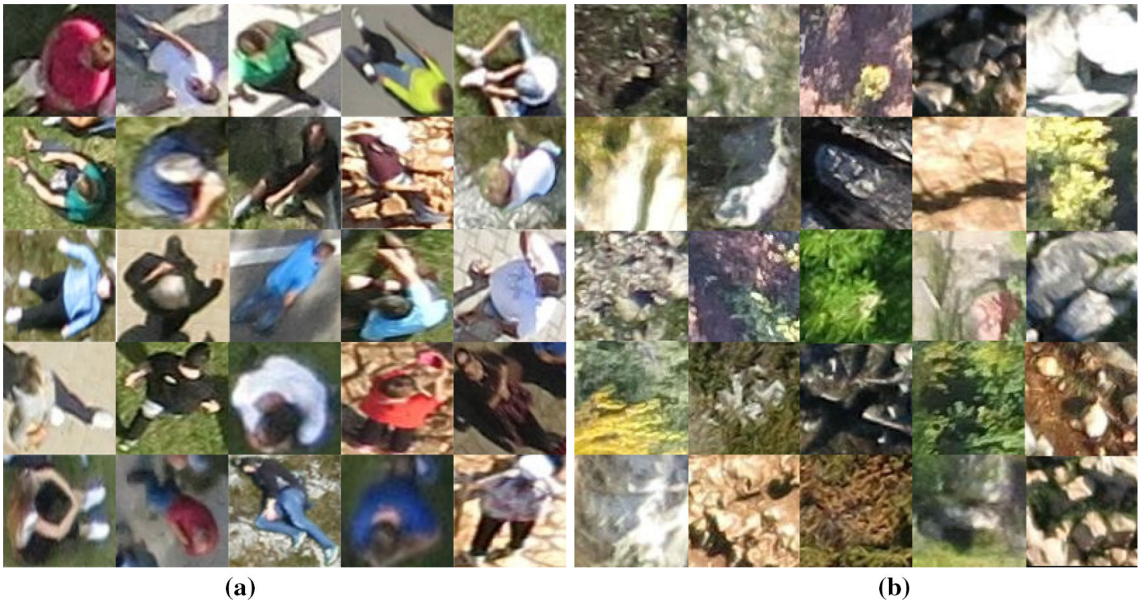


Fig. 8 Training dataset excerpt: **a** positive samples (people), **b** negative samples (background)

iteration, the threshold value is increased at a step size of 5. On average, it is necessary to perform several iterations to reduce the number of blobs to the target value of 300 or fewer blobs per saliency map. After producing the binary map and extracting the blobs, we further filtered some blobs based on size and a priori knowledge of the dimensions of the target objects. This prior knowledge is used as a preprocessing step to eliminate large areas or regions that we presumably know are not people (e.g., roads, forests, meadows) or small areas or blobs that occupy less than 12 × 12 pixels. To merge close blobs, we dilated the entire image using a kernel size of 7. By applying this procedure to all 101 images, we extracted 15,048 blobs, or approximately 14,899 blobs per image. This number represents a significant reduction of the potential candidates. For the remaining parts of the binary map, we calculated the centroids or (x, y) central coordinates of each individual blob. The coordinates of a blob on the binary map correspond to its coordinates in the RGB blob representation in the original image because the saliency map, binary map and original image all have the same size. In this way, it is relatively easy to isolate parts of the colour image or patches without additional computations and forward them to the CNN module for classification. An illustration of the region selection described above is shown in Fig. 9.

Figure 9a shows the colour RGB image from a remote location. Figure 9b, c represent the corresponding saliency and binary maps, respectively, and Fig. 9d represents the filtered and selected RGB regions. As we can see from Fig. 9d, this approach significantly reduces the search space area and the number of region proposals. Table 5 shows the performance of the region proposal module per location on our test dataset of images.

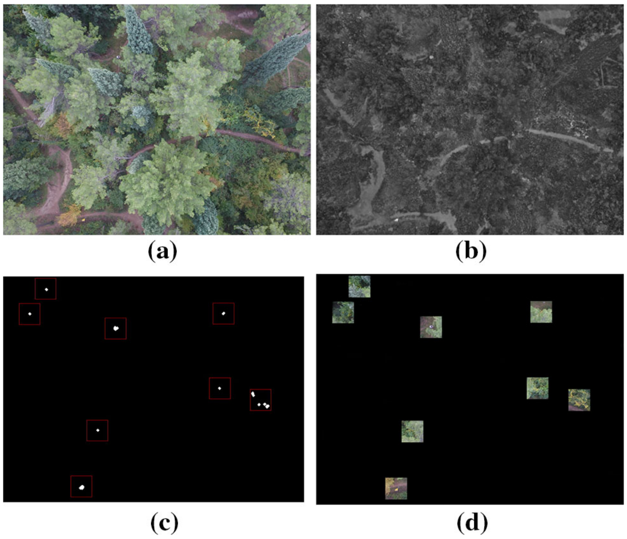


Fig. 9 Region proposal procedure: **a** GB image. **b** Saliency map. **c** Binary map. **d** Selected regions

From a total of 341 ground-truth (GT) lost persons in 101 images, RPM successfully detected 317 region proposals that include a person or part of a person, achieving a detection rate of 93.0% compared to the ground truth. Evaluating all 148.99 region proposals per image through a visual inspection would be a tedious task for a human observer; in fact, it would actually be quite counterproductive. The number of region proposals needs to be drastically reduced in the next stage. Overall, 15,048 selected salient blobs, a mean of 148.9 per image, are forwarded to the classification module for further inspection. This total includes 317 blobs that represent a person or part of a person. At this stage, the model failed to

Table 5 Region proposal module performance on test images

Location	Number of images	Ground truth	Recall	Number of ROIs	ROIs per location
BLI	9	47	100.0%	1672	1858
BRA	12	22	86.4%	2006	1672
BRK	12	19	100.0%	1561	1301
BRS	11	21	95.2%	1685	1532
CAB	8	34	88.2%	1292	1615
CAP	6	19	100.0%	909	1515
MED	6	53	81.1%	864	1440
MOB	5	13	100.0%	161	322
SB	10	26	100.0%	1088	1088
TRS	6	15	100.0%	672	1120
ZRI	6	49	89.8%	1120	1867
GOR	10	20	100%	2018	2018
Overall	101	341	93.0%	15,048	14,899

detect 24 ground-truth blobs; consequently, those were not available for the subsequent classification stage.

5.2 CNN Classification Module

With sufficient data, deep CNNs in most cases give state-of-the art results, especially for classifying objects in images. However, as mentioned in the previous chapter, no public training set of people in natural environments has been gathered from UAV images, and in most cases, relatively sparse training datasets lead to an overfitting problem. In most cases, to resolve this issue, we need to acquire more data and/or use transfer learning by using models pre-trained on some other, larger image dataset. These pre-trained models can be adapted to specific tasks either by fine tuning (using the network parameters as initialization and re-training with the new dataset) or by using them as simple feature extractors for the recognition task. Determining whether to use fine-tuning or feature extraction depends on two main factors: the size of the new dataset and its similarity to the original dataset. Fine-tuning is recommended when we have many data and the new dataset is similar to the original dataset. Additionally, it is considered beneficial to fine-tune through the entire network if we have a large dataset that is different from the original dataset (although in this case we could also simply train a network from scratch). We decided to use fine-tuning as a transfer-learning method because we managed to acquire a relatively modest dataset of person patches for training positive "person-like" as well as negative background "no-person" patches. Positive samples were labelled on the training images that contained persons.

Before fine-tuning the pre-trained VGG-16 network on ILSVRC (Russakovsky et al. 2014), the patches were first resized to 224 × 224 pixels and then zero-cantered by mean pixel from the ImageNet dataset. Training was conducted using the Caffe deep learning framework on a workstation equipped with an Nvidia GTX Titan X graphic card.

5.3 Results and Comparative Study

After all the images from the testing collection were processed by the RPM module described in Sect. 5.1, we obtained a corresponding binary map for each image. Binary maps consist of blobs or object suggestions. These object suggestions represent anchors around which we form regions or patches that are used as input to the VGG16 network or CNN module. Note that in the majority of cases, the anchors do not segment the entire target object; they cover only some part of the person's body such as the feet, shoulders, head, or other body part. Consequently, we need to consider the context and define an area around anchors that is sufficiently large to extract enough features to describe a person. For every anchor, we use three scales (1:1, 1:1.16, 1.16:1). Although the test images were obtained from comparable altitudes, due to large variations in possible person poses and camera characteristics from the different UAV platforms, it was not easy to ascertain an appropriate size for the bounding boxes. After experimenting with a variety of sizes, we concluded that the best starting size for the bounding box is approximately 60 × 60 pixels. On the centroid of each anchor, we formed a region using a certain window size and cropped it. To acquire images for input to the VGG16 network, each cropped region was scaled and normalized to 224 × 224 pixels. After the images were passed through the convolutional and other layers in the network, we obtained a score based on the SoftMax classifier. Based on this score, the regions were classified as either "person-like" or "not person" (background). The following table and graphs show the results from using various selected window sizes on the anchors.

From Table 6 and the corresponding graphs it can be seen that accuracy and recall decrease with respect to window

Table 6 Analysis in regard to the anchor window size (Ground Truth = 341; ROI = 15,048)

Window size in pixels	Detected objects (TP + FP)	TP	FP	FN	Recall (%)	Precision (%)	Accuracy (%)
52 × 52	1203	299	904	42	87.7	24.9	93.7
56 × 56	867	297	570	44	87.1	34.3	95.9
60 × 60	750	296	454	45	86.8	39.5	96.7
64 × 64	629	282	347	59	82.7	44.8	97.3
68 × 68	532	277	255	64	81.2	52.1	97.9

size, while the opposite is true for precision. Using a window size of 52 × 52 pixels, the model detected 1203 objects, of which 299 were true positives and 904 were false positives. The recall was 87.7%, and the precision was 24.9%. In contrast, when using a window size of 68 × 68 pixels, the model detected considerably fewer objects (532), of which 277 were true positives and 255 were false positives. The recall was 81.2%, and the precision was 52.1%. These numbers represent a considerable reduction in recall, but the precision more than doubled. Based on these observations, we decided to adopt a window size of 60 × 60 pixels as our referent window size. Using this referent widow size, the model achieved a recall of 86.8%: only 4 fewer true positives than with a window size of 52 × 52 pixels. However, it resulted in only 450 false positives, which improved the precision to 39.5%. Accuracy is a measure of system performance with regard to both correctly detecting and correctly rejecting targets. Using a window size of 60 × 60, the accuracy was 96.7%. To further enhance the detection model for each image, three most prominent points were selected and included in such a way that they completely skipped the CNN classification module. By combining and incorporating the CNN module and the TOP3 salient points from the RPM module, we obtained the final results for our proposed system, which are shown in Table 6. Using this approach, we managed to increase the true positives by 7, which led to an increase in recall of 2.1% (88.9%). The number of false positives increased by 114, leading to a 4.7% drop in precision (34.8%) rate compared to using only the CNN. Nevertheless, the precision score is still better than when using smaller-sized windows (56 × 56 or 52 × 52) and results in a much better detection rate.

5.3.1 Comparative analysis with the method based on the mean-shift algorithm and Faster R-CNN

For comparative analysis, we included a detection model from a previous research paper within the IPSAR project as well as state of the art Faster R-CNN. The model within the IPSAR project is based on the mean-shift segmentation algorithm (Turić et al. 2010; Music et al. 2016) model which mostly uses heuristic rules such as the size of segments and clusters for decision making. The authors chose the mean-shift algorithm for segmentation mostly because it demonstrated good results regarding segmentation quality and stability. To reduce the high computational requirements and the quadratic computational complexity of the algorithm, they decided to modify and use two-stage mean-shift segmentation, which caused only minor loss of accuracy.

Faster R-CNN has shown excellent results in many complex computer vision detection problems. It is a purely CNN-based two-stage architecture consisting of two modules (a region proposal network module and a Fast RCNN detector) (Ren et al. 2015). Its main advantage is that both modules can share the bottom convolutional layers for the whole image. In this way, the RPN can use deep neural networks such as VGG16 to generate high-quality proposals, making the entire process of proposal generation almost computationally cost-free. The Faster R-CNN detector takes multiple regions of interest (ROIs) from a shared convolutional feature map as input. In the RoI pooling layer, a fixed-length feature vector is extracted for every RoI and propagated through a sequence of fully connected (FC) layers. Finally, it outputs branches into two sibling output layers: one produces SoftMax probability estimates over K object classes plus a catch-all background class, and the other layer outputs four real-valued numbers for each of the K object classes.

Faster RCNN model is not trained on image patches but on labeled images. Therefore, the images described in Sect. 4 are used for training purposes. In training there is no need for explicit allocation of negative samples because the negative or background class is calculated automatically by the FasterR-CNN framework. We used a stochastic gradient descent (SGD) solver with 40 K iterations and a learning rate of 0.001,which are the same values used to train our proposed model.In its original form, the Faster RCNN detection framework is not well-optimized for training and testing models in large format images. Mainly due to memory requirements for the GPU framework it is not possible to process our image dataset at original resolution; Instead, it reduces or scale them to a feasible resolution. We experimentally determined that the maximum resolution supported by the test GPU on the VGG16 network is one third of each axis (1333 × 1000 px). Scaling does not pose a particular issue in detection appli-

Table 7 Comparison of models on IPSAR database of images

Location	GT	Mean-Shift model (Turić et al. 2010)		Faster RCNN (Ren et al. 2015)		Our model	
		Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)
BLI	47	74.5	36.8	87.2	71.9	97.9	42.6
BRA	22	81.8	14.9	72.7	39.0	86.4	27.5
BRK	21	76.2	12.8	90.5	65.5	100	28.0
BRS	21	61.9	14.3	76.2	23.2	85.7	27.7
CAB	34	73.5	29.1	85.3	72.5	82.4	45.9
CAP	19	89.5	21.3	94.7	81.8	100.0	51.4
MED	53	66.0	27.8	84.9	71.4	75.7	35.5
MOB	13	100.0	17.8	100.0	54.2	100.0	50.0
SB	24	73.1	12.0	79.2	57.6	87.5	42.0
TRS	18	61.1	10.7	66.7	52.2	94.4	18.3
GRO	20	90.0	11.8	90.0	54.6	90.0	28.1
ZRI	49	71.4	22.3	89.8	67.7	87.8	39.1
Overall	341	74.7	18.7	85.0	58.1	88.9	34.8

cations where the target object is relatively large, but in applications where the target objects are small, it significantly undermines the detection performance. The one-third scaling operation reduces the spatial information by a factor of 9, which consequently leads to extremely small person blob objects that are unable to meet threshold parameters or anchor settings in the RPN network or in the Faster R-CNN detector. Therefore, instead of rescaling, we decided to crop the original images in segments of 1333 × 1000 px (with an overlapping offset).

Table 7 presents the results of the two-stage mean-shift-based model from Turić et al. (2010), Faster R-CNN (Ren et al. 2015) and our presented model.

As Table 7 shows, our proposed model achieved better results compared to the model from the previous research. Our proposed model achieved an 88.9% detection rate, an improvement of 14.2%. In absolute amounts, out of a total of 341 searched objects in the set, our model successfully recognized 303 objects, an improvement of 48 persons. From the graph, at the ten locations, our model achieved a significant improvement in recall. In absolute terms, the largest improvement was achieved at the BLI location, where the proposed model had 11 more hits, while in relative terms, the greatest improvement occurred at the BRK location, showing a difference of 34.8% in favour of our model. In two locations, MOB and GOR, the system based on mean-shift segmentation achieved the same results as our system. Regarding precision, our model achieved better results at all locations by significant margins. The greatest improvement in both absolute and relative terms was at location SB, where our model yielded 110 fewer false alarms or a 30% better precision rate. Overall, our model achieved a precision of 34.8%, which is 16.1% better than the previous model.

As shown in Table 7, our proposed model achieved an average 3.9% improvement compared to Faster RCNN detection framework, which is important for SAR missions. However, Faster RCNN achieved better precision than our model, mainly because it uses a different mechanism for negative background sample generation. For every positive ground truth sample, it generates three hard negative background samples by choosing ambiguous, confusing samples that are harder to correctly classify. By default, approximation of hard negative samples is done by choosing samples that slightly overlap with ground truth positives (IoU < 0.5). This mechanism and fact that Faster RCNN uses a larger number of negative background samples (three negatives to one positive) contributes to better overall precision. We could improve the precision of our model by increasing the number of negative samples number.

It can be seen from the graph - that our model achieved better results in recall, at the seven locations. In absolute terms, the largest improvement was achieved at the BLI and TRS location, where the proposed model had 5 more hits, while in relative terms, the greatest improvement occurred at the TRS location, showing a difference of 27.3% in favour of our model. At three locations, ZRI, CAB and MED, the system based on Faster RCNN achieved better results and this is particularly emphasized at MED location. Regarding precision, Faster RCNN model achieved better results at all locations by significant margins which indicates that it is more tuned to precision than recall.

Additional comparison of these methods is presented in Figs. 10, 11, 12, 13 and 14. In these figures our system detected all the "lost" persons, while the mean-shift-based model had problems detecting persons dressed in black and white clothing or clothing similar to the background colours.

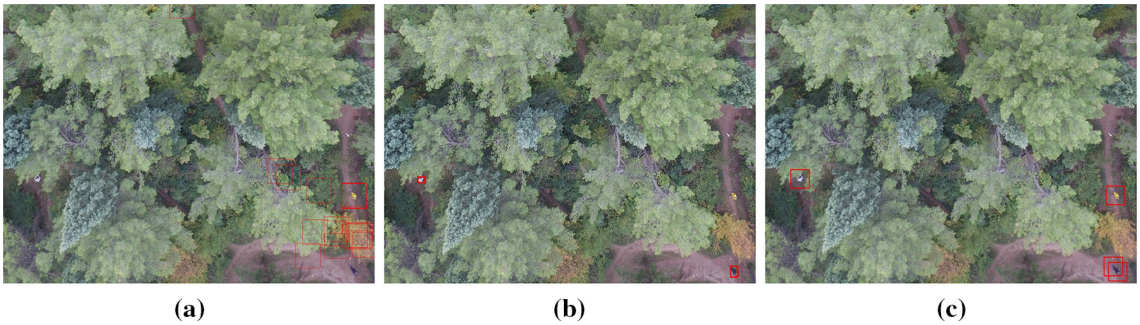


Fig. 10 Image at TRS location: **a** IPSAR system. **b** Faster R-CNN. **c** Proposed system

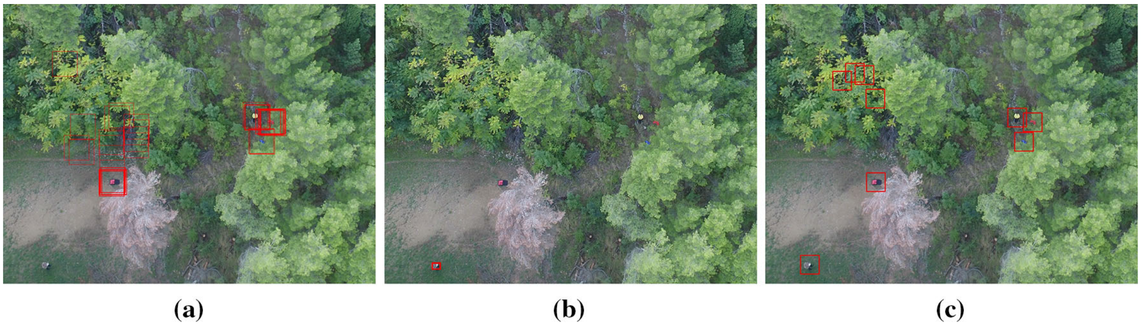


Fig. 11 Image at BRK location: **a** IPSAR system. **b** Faster R-CNN. **c** Proposed system

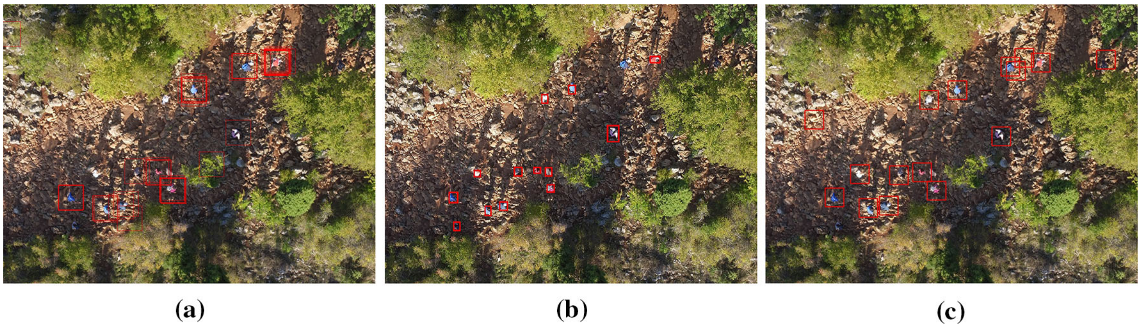


Fig. 12 Image at MED location: **a** IPSAR system. **b** Faster R-CNN. **c** Proposed system

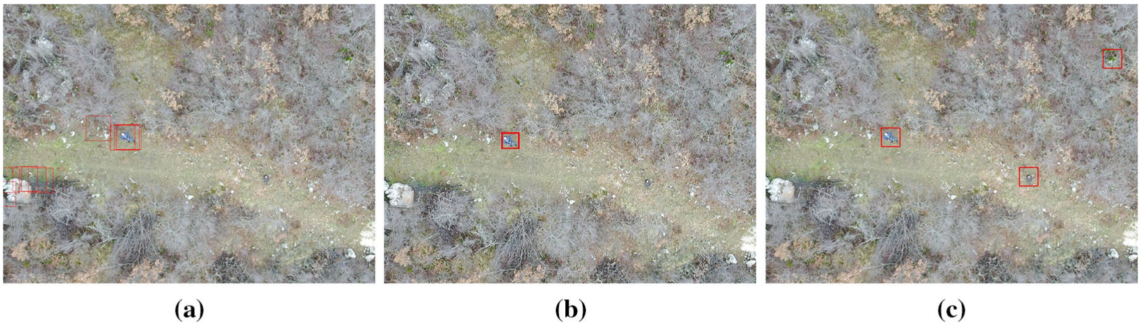


Fig. 13 Image at TRS location: **a** IPSAR system. **b** Faster R-CNN. **c** Proposed system

Faster RCNN had problems with detecting people in yellow and dark clothing. Although this study is primarily interested in detecting people, Faster RCNN was not able to detect items such as bags, jackets and similar items that can function as clues of human presence unlike the other two models. Figure 12 is particularly interesting because it represents a rather complex scene that includes vegetation and rocky environments as well as many objects of interest. The complexity of

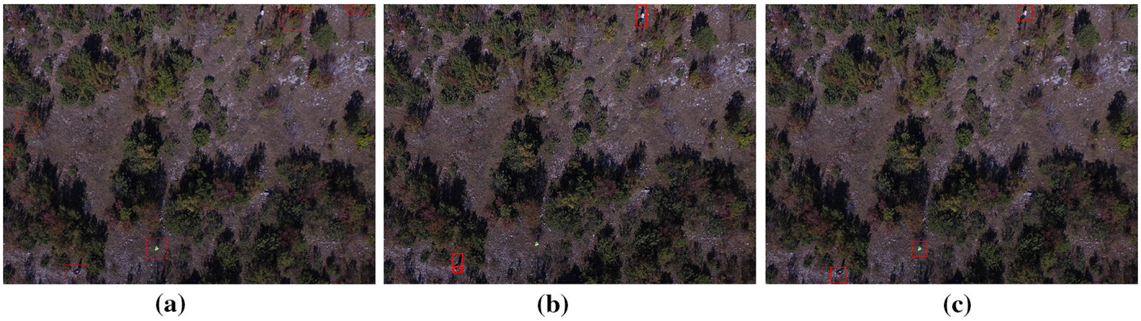


Fig. 14 Image at BRS location: **a** IPSAR system. **b** Faster R-CNN. **c** Proposed system

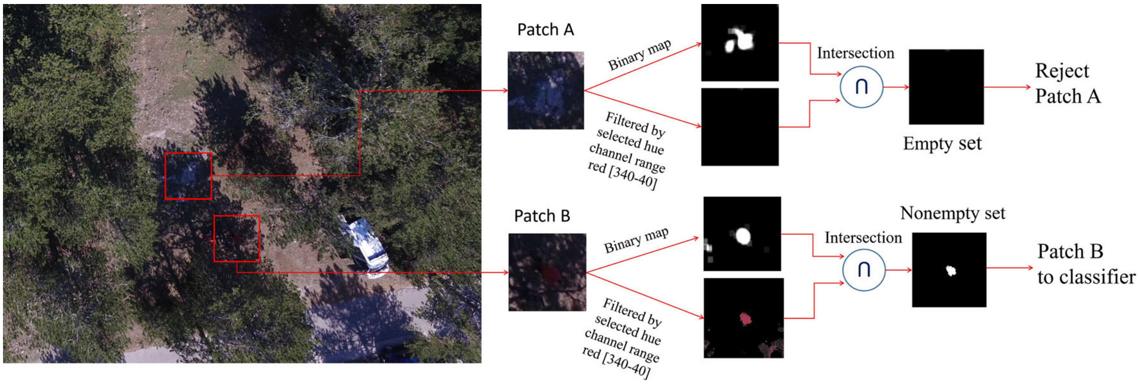


Fig. 15 Using colour information to select additional proposed regions (Color figure online)

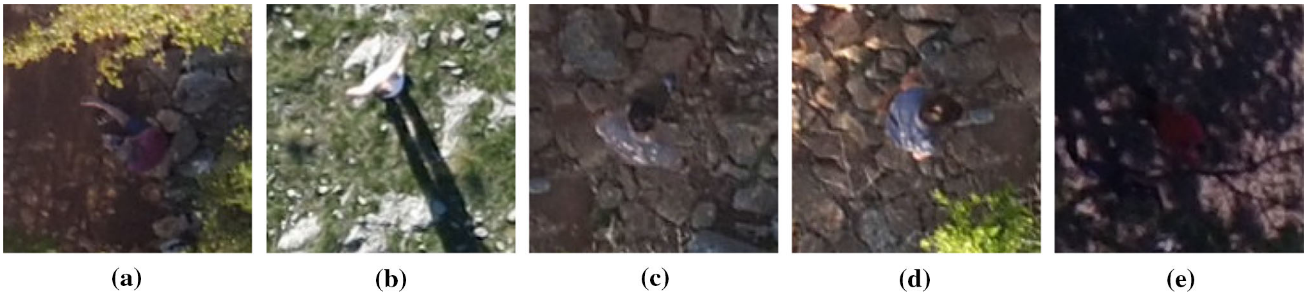


Fig. 16 Selected patches missed by RPM module

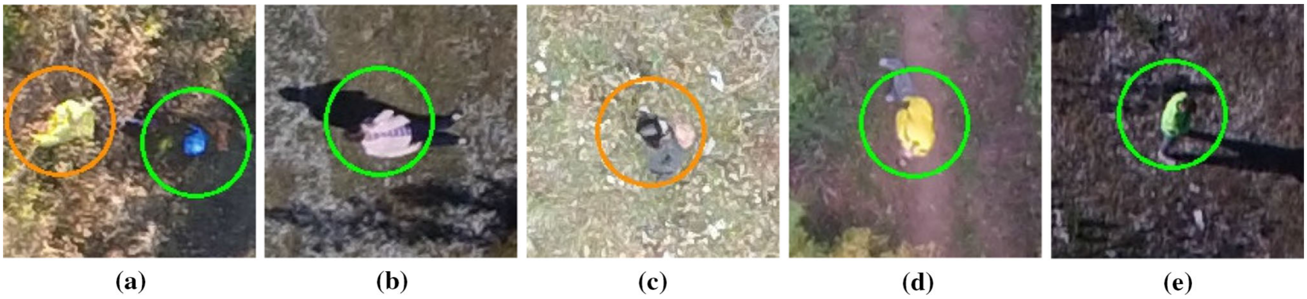


Fig. 17 Selected patches missed by CNN module. Green circles are related to those "missing" persons that the CNN did not detect but which were detected by the RPM module. Orange circles represent "missing" persons that were not detected, either by the CNN module or by the RPM (Color figure online)

the scene is further aggravated by the sun, which results in reflections and shadows. These are some of the reasons why all three models failed to detect all target objects. Our proposed model and Faster RCNN detected 13 out of 16 objects of interest, while the mean-shift-based model detected only 11. In the selected images, as well as in all others, our model detected more target objects, while Faster RCNN had fewer false alarms. Our model has another advantage specific to SAR missions compared to the Faster R-CNN. At the beginning of a SAR mission, planners gather as much relevant information about missing persons as possible. This includes information about clothing colours, which can significantly contribute to a more successful search. Therefore, we included this as one possibility in our model. When the sought person's clothing colour is known, it can be specified as additional search criteria. In this case, our region proposal module processes an additional 200 salient patches. These patches are converted to HSV colour space, over-segmented by the SLIC algorithm and propagated to a colour filter. If this patch responds to the colour range, the patch is forwarded to the CNN and classified. This process is illustrated in Fig. 15. Using this approach resulted in an increase in the number of positively detected people.

6 Discussion

From Table 5 in subsection 4.1, it is evident that the region proposal module did not detect all the possible candidate regions with perfect accuracy. The overall recall was 93.0%: in absolute terms, it detected 317 regions correctly but missed 24 true candidate proposals. In four locations (BRA 86.4%, CAB 88.2%, MED 81.1% and ZRI 89.8%), we obtained below average results. Some of the missed persons in these locations are shown in Fig. 16. The RPM based on wavelet decomposition failed to detect the persons shown in Fig. 16a–e. mostly due to the relatively weak local contrast in the surrounding area. This is particularly apparent in Fig. 16a, c, e. The unusual contours of a person in a white t-shirt blended well with the environment in Fig. 16b and make it very hard for even a human observer to recognize. The mountain location CAB, which includes many interspersed rocks makes the target in Fig. 16b difficult for the RPM module to identify as salient. In this scene, the error occurred mainly due to the global contrast computation in the salient detection algorithm.

A loss of 24 true candidates or 7.0% is significant because it cannot be reversed in a later stage by the CNN module, which in fact demonstrated very good results. From the total 317 true positives forwarded by the RPM module, 296 proposals were correctly classified as "person", resulting in a classifier recall of 93.4%. In Fig. 17, we selected some patches that were incorrectly labelled as background by the CNN module. Individual images are marked with green or orange circles around the objects of interest. The green circles are related to those "missing" persons that the CNN did not detect but which were detected by the RPM module using the TOP3 rule. This rule, as discussed earlier, detects three most prominent objects unconditionally and labels them as a positive class. This approach introduces some amount of uncertainty in the whole system because while it can cause the system to correctly label a positive class, as shown in Fig. 17b, d, e it can also result in some objects (blue bag) being mislabelled as positive, as shown in Fig. 17a. The orange circles represent "missing" persons that were not detected, either by the CNN module or by the RPM TOP3 rule. Figure 17a shows a person in a yellow jacket in a highly occluded environment, so it is understandable that that object was wrongly classified as background. Nevertheless, the system managed to detect the nearby blue bag; thus, it is highly possible for a human observer to spot a person by using this nearby clue during the visual inspection process. Objects such as bags, jackets and other clues that lost people can leave behind can also contribute to a successful find; however, our CNN module is not well trained to detect such objects. Therefore, it is highly beneficial to include certain heuristic rules, expert knowledge or a priori information to make the whole system more robust. These are some of the main reasons why we decided that a certain number (or more precisely, 3) of most significant salient points should be selected without being further processed in the CNN module. Using this approach, we managed to significantly improve the detection results (recall).

7 Conclusion

In this paper, we proposed a region-based CNN person detection framework to support ground search and rescue missions. Also, we compile public available image database of UAV imagery of lost or missing persons in natural environments called HERIDAL. The main focus in this study is on supporting missions that occur in remote, wild or non-urban areas where UAVs are used for gathering terrain images that are later inspected by our computer vision model. The model's goal is to detect and suggest possible locations where people might be present in images. The model consists of two separate parts. In the first part, or region proposal module, we modified the class-agnostic salient detection algorithm to filter and reduce the search space and to propose blobs as building blocks for patch generation. The patches are then further processed in the CNN classification module. The CNN module is based on the VGG16 architecture, and its role is to classify the proposed RoI into one of two classes: person or non-person. The experiments and testing were conducted on a HERIDAL image dataset collected in a variety of weather conditions and at various locations in Croatia, Bosnia and Herzegovina. HERIDAL comprises over 68,750 wilderness image patches taken from an aerial perspective intended for training, as well as approximately 500 labelled full-size real-world images intended for testing. Our proposed model was inspired by RCNN methods, and it has achieved significantly better results in detecting the presence of people in images, surpassing previous methods based on segmentation and heuristic decision-making that are used by Croatian SAR teams. The overall detection rate achieved by our proposed method was 88.9%, while the false alarm rate was 34.8%. We compared the proposed model with the state-of-the-art Faster R-CNN, trained on the HERIDAL database, and our proposed model showed better performance for SAR applications. In future work, we plan to expand the image database and make it even more relevant, thus further encouraging research in this important area. The demonstrated effectiveness of deep learning methods and the overall proposed system can still be improved, especially with regard to the false alarm rate and the execution speed. In this paper, development of a real-time system was not required; however, this real-time performance is an aspect that we plan to address in future research.

Acknowledgements This research was carried out in part within the framework of a IPSAR project, University of Split, Croatia. It is also partly supported by Federal Ministry of Education and Science, Bosnia and Herzegovina by Grant (NG 05-39-2945-3/16) to Faculty of Science and Education, University of Mostar. We thank NVIDIA Corporation for GPUs donation through Nvidia GPU Edcuation Center program at University of Mostar.

Appendix A

Algorithm 1: Two-stage segmentation of aerial image for SAR pseudocode

```
Data: Color RGB images
Result: Segmented regions potentially with human presence
transform image to YCbCr color space;
apply median filter to Cb and Cr components;
divide image into subimages;
foreach subimage do
    run mean shift clustering algorithm;
    append K with subimage cluster matrix;
end
apply mean shift to global cluster matrix K;
return set of resulting clusters;
foreach cluster do
    if $cluster\_size \geq N_{MAX}$ then
        ignore cluster;
    end
    find set of spatially connected areas;
    foreach area do
        if $area\_size \geq N_{MIN}$ then
            eliminate area;
        end
        dilate image to merge nearby segments;
    end
    if $q \geq N_A$ then
        ignore cluster;
    end
    outline potential target regions
end
```

References

Angelova, A., Krizhevsky, A., Vanhoucke, V., Ogale, A., & Ferguson, D. (2015). Real-time pedestrian detection with deep network cascades. In *Proceedings of BMVC 2015*.

Anna Gaszczak, J. H., & Breckon, Toby P. (2011). *Real-time people and vehicle detection from UAV imagery* (Vol. 7878, pp. 7878–7878-13). https://doi.org/10.1117/12.876663.

Borji, A., Cheng, M. M., Hou, Q., Jiang, H., & Li, J. (2014). *Salient object detection: A survey*. arXiv preprint arXiv:1411.5878.

Chen, C., Liu, M. Y., Tuzel, O., & Xiao, J. (2017). R-CNN for small object detection. In S. H. Lai, V. Lepetit, K. Nishino, & Y. Sato (Eds.), *Computer Vision—ACCV 2016* (pp. 214–230). Cham: Springer.

Daubechies, I. (1992). *Ten lectures on wavelets*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Eggert, C., Brehm, S., Winschel, A., Zecha, D., & Lienhart, R. (2017). A closer look: Small object detection in faster R-CNN. In *2017 IEEE international conference on multimedia and expo (ICME)* (pp. 421–426). https://doi.org/10.1109/ICME.2017.8019550.

Enzweiler, M., & Gavrila, D. M. (2009). Monocular pedestrian detection: Survey and experiments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(12), 2179–2195. https://doi.org/10.1109/TPAMI.2008.260.

Girshick, R. B. (2015). *Fast R-CNN*. CoRR arXiv:1504.08083.

Girshick, R. B., Donahue, J., Darrell, T., & Malik, J. (2013). *Rich feature hierarchies for accurate object detection and semantic segmentation*. CoRR arXiv:1311.2524.

Gotovac, S., Papić, V., & Marušić, Ž. (2016). Analysis of saliency object detection algorithms for search and rescue operations. In *24th International conference on software, telecommunications and computer networks (SoftCOM)* (pp. 1–6). https://doi.org/10.1109/SOFTCOM.2016.7772118.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep residual learning for image recognition*. CoRR arXiv:1512.03385.

Hosang, J., Omran, M., Benenson, R., & Schiele, B. (2015). Taking a deeper look at pedestrians. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Imamoglu, N., Lin, W., & Fang, Y. (2013). A saliency detection model using low-level features based on wavelet transform. *IEEE Transactions on Multimedia*, *15*(1), 96–105. https://doi.org/10.1109/TMM.2012.2225034.

Koch, C., & Ullman, S. (1987). *Shifts in selective visual attention: Towards the underlying neural circuitry* (pp. 115–141). Dordrecht: Springer. https://doi.org/10.1007/978-94-009-3833-55.

Koester, R. (2008). *Lost person behavior: A search and rescue guide on where to look for land, air, and water*. dbS Productions. https://books.google.hr/books?id=YQeSIAAACAAJ.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th international conference on neural information processing systems—Volume 1, Curran Associates Inc., USA, NIPS'12* (pp. 1097–1105). http://dl.acm.org/citation.cfm?id=2999134.2999257.

Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. In *Proceedings of the IEEE* (pp. 2278–2324).

Leroy, J., Riche, N., Mancas, M., Gosselin, B., & Dutoit, T. (2014). *Superrare: an object-oriented saliency algorithm based on superpixels rarity*.

Li, J., Levine, M. D., An, X., Xu, X., & He, H. (2016). *Visual saliency based on scale-space analysis in the frequency domain*. CoRR arXiv:1605.01999.

Music, J., Orović, I., Marasović, T., Papić, V., & Stanković, S. (2016). Gradient compressive sensing for image data reduction in UAV based search and rescue in the wild. In *Mathematical problems in engineering, 2016*. https://doi.org/10.1155/2016/6827414.

Ren, S., He, K., Girshick, R. B., & Sun, J. (2015). *Faster R-CNN: Towards real-time object detection with region proposal networks*. CoRR arXiv:1506.01497.

Rudol, P., & Doherty, P. (2008). Human body detection and geolocalization for UAV search and rescue missions using color and thermal imagery. In *2008 IEEE aerospace conference* (pp. 1–8). https://doi.org/10.1109/AERO.2008.4526559.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2014). *Imagenet large scale visual recognition challenge*. CoRR arXiv:1409.0575.

Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition*. CoRR arXiv:1409.1556.

Sokalski, J., Breckon, T. P., & Cowling, I. (2010). Automatic salient object detection in uav imagery. In *Proceedings of the 25th international unmanned air vehicle systems* (pp. 1–12).

Syrotuck, W., & Syrotuck, J. (2000). *Analysis of lost person behavior: An aid to search planning*. Barkleigh Productions. https://books.google.hr/books?id=3rWDAAAACAAJ.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In *2015 IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 1–9). https://doi.org/10.1109/CVPR.2015.7298594.

Tian, Y., Luo, P., Wang, X., & Tang, X. (2015). Deep learning strong parts for pedestrian detection. In: *2015 IEEE international conference on computer vision (ICCV)* (pp. 1904–1912).

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognit Psychol*, *12*(1), 97–136.

Turić, H., Dujmić, H., & Papić, V. (2010). Two-stage segmentation of aerial images for search and rescue. *Information Technology and Control*, *39*, 138–145.

Viola, P., Jones, M. J., & Snow, D. (2003). Detecting pedestrians using patterns of motion and appearance. In *Proceedings ninth IEEE international conference on computer vision* (Vol. 2, pp. 734–741). https://doi.org/10.1109/ICCV.2003.1238422.

Yuan, P., Zhong, Y., & Yuan, Y. (2017). *Faster r-cnn with region proposal refinement*.

Zendel, O., Murschitz, M., Humenberger, M., & Herzner, W. (2017). How good is my test data? Introducing safety analysis for computer vision. *International Journal of Computer Vision*, *125*(1–3), 95–109. https://doi.org/10.1007/s11263-017-1020-z.

Zhang, L., Lin, L., Liang, X., & He, K. (2016). Is faster R-CNN doing well for pedestrian detection? In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer Vision—ECCV 2016* (pp. 443–457). Cham: Springer.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.