POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Anomaly Detection from Aerial Images for Search and Rescue Missions

Laurea Magistrale in Information Engineering - Ingegneria Informatica

Author: Luca Morandini

Advisor: Prof. Piero Fraternali

Co-advisor: Federico Milani

Academic year: 2021-2022

1. Introduction

Search and Rescue (SAR) missions are time-critical and challenging operations. The organization of missions in inaccessible areas is a complex task that requires scanning vast terrains, especially when the location of the missing people is imprecise or unknown.
Unmanned Aerial Vehicles (UAVs) equipped with optical or thermal cameras can quickly scan the area of an accident and collect a massive amount of images to identify the locations of the missing people in a short amount of time. However, the large quantity of captured images must be manually scanned to identify people or objects of interest, such as clothing or technical equipment.
Computer Vision (CV) techniques, combined with Machine Learning (ML) or Deep Learning (DL) models, can speed up this process by filtering frames and highlighting signals of human presence, reducing the burden of manually screening the entire captured video sequence.
This work approaches the people detection problem from thermal drone imagery in SAR missions as an anomaly detection task. Thermal sensors can measure the temperature difference between human bodies and the background, even in dense forests where the trees block sunlight. Due to the nature of the task, background images (i.e., those images without targets) are abundant, while very few frames contain people or objects of interest. For this reason, the implemented ML and DL models are trained to learn a background representation and should signal the presence of anomalies, i.e., all those areas that do not belong to a typical known background scene.
The chosen methods are evaluated quantitatively and qualitatively. The best ML technique can obtain 73.5% F1-Score on the anomaly class, while the best DL model presents an outstanding 92.6% F1-Score. To explain the model results, anomaly heatmaps are generated on frames with and without targets.
The developed system is designed to collaborate with the rescue teams by filtering most of the background images to: 1) reduce the burden of manually screening the entire captured videos and 2) highlight areas of the images with higher chances of containing an injured or lost person.

2. Related Work

Anomaly detection is the set of techniques used to identify patterns in data that deviate from expected behavior. A common approach is to model the representation of the normal data and

any observation that does not belong to the normality distribution is considered an anomaly. Therefore, it is essential to develop a model capable of identifying anomalous behavior only by leveraging normal data.
Anomaly detection on image data requires an additional step for extracting valuable features that characterize the image, due to the high dimensionality of data and the locality property of pixels information. Texture is a very useful characterization for a wide range of images. In fact, a large number of texture feature extraction methods are proposed in the literature. Feature extractors such as Haralick features, Local Binary Pattern (LBP), and Histogram of Gradient Magnitudes (HGM) are statistical methods based on the analysis of the spatial distribution of gray-level values. Particularly, Haralick and LBP are based on the study of the gray levels in the neighborhood of the pixel whereas HGM derives a histogram of the gradient magnitudes computed on the entire image. Other structural approaches decompose the image into primitives and their spatial arrangements are used to characterize textures. For instance, techniques such as Gabor Decomposition, Wavelet Transform, and LETRIST apply filters on an image to derive features that are combined to generate the final descriptor of the image texture.
The features extracted from an image are combined into a *feature vector*, then classified with an anomaly detection model. Some classification techniques such as One-Class SVM and SVDD learn a boundary to separate the normal class from the anomalies in the feature space. Other methods such as Isolation Forest explicitly separate each sample from the others, identifying as anomalies the points that can be isolated in fewer steps. Differently, Local Outlier Factor identifies the anomalies as points characterized by a lower local density with respect to the densities of their neighbors.
DL techniques have more effective anomaly detection capabilities due to their ability to fine-tune the representation of normal data. Consequently, they can better identify the anomalies that diverge from the learned distribution. Autoencoders, such as DCAE and RCAE, and Generative Adversarial Networks, such as AnoGAN or GANomaly, learn to reconstruct normal images. Thus, they identify anomaly samples by evaluating the reconstruction error that should be higher for anomalous instances. Other techniques such as OC-NN, Deep SVDD, and Deep SAD learn a representation of normality by mapping normal samples to an area in latent space enclosed by a boundary. Points mapped outside of the boundary are then classified as anomalies.
In the literature, few data sets are dedicated to the Search and Rescue scenario. The only available data set captured over forest scenarios is *Data: Search and Rescue with Airborne Optical Sectioning* [5]. The data set comprises 12 flights performed at an altitude of 30-35m over different types of forest (broadleaf, conifer, mixed) and 6 flights in open field. For each flight, thermal and RGB frames are provided, aligned to cover the same view of the scene. The ground truth comprises 9,684 annotated bounding boxes that enclose the entire body of all the people. However, due to the occlusions caused by the trees, a person could be partially or entirely hidden.
Some applications that assist SAR missions exploit spectral information of drone imagery to identify color anomalies that may indicate the presence of a person. These systems do not search for specific patterns in the image, therefore they may generate a high False Positive Rate. Most recent works, approach the problem of people localization in rescue missions using object detection models trained on data sets that specifically simulate SAR scenarios [4]. However, since these models learn to identify the shapes and textures of people, they could miss some targets that are partially occluded. To improve ground visibility, some applications rely on thermal sensors which are less influenced by occlusions. Thermal images better identify human bodies that have warmer temperatures than the surrounding background.

3. Data Set and Methods

The data set presented in [5] was originally designed for fully supervised object detection tasks and all the images, from each flight, were merged to remove tree occlusions and make the targets more visible. This data set has been adapted to be used in this work which aims at developing an anomaly detection system to quickly locate potential targets without requiring accurate pixel-level localization. In fact, the anomaly detection problem has been approached as a classification

task of tiles extracted from images captured during a mission. Each generated tile is classified as *background* or *anomaly* depending on whether or not a target is detected on the image. In our case, only *background* tiles are used for training, while all tiles (*background* and *anomaly*) are employed for evaluation. Thus, new training, validation, and testing splits needed to be defined.
The images of the available data set [5] captured over various forest types are visually different and have distinctive characteristics such as the shape of the trees and the degree of ground occlusion, as shown in Figure 1. To avoid overfitting the features of a forest type, the training data should include samples from flights captured on different forests.

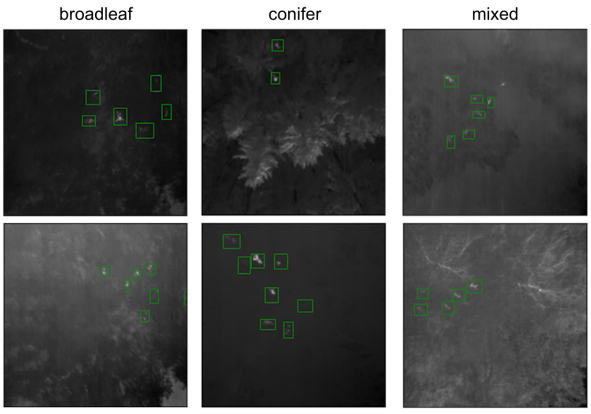

Figure 1: Images from different forest types.

The split was performed to distribute 70% of the annotations in the *training set* with images from broadleaf, conifer and mixed forests, 10% of the annotations from a mixed forest flight in the *validation set*, and 20% of the annotations in the *test set* which are captured over broadleaf and conifer forests.
After researching the state-of-the-art feature extraction techniques [1] and defining desired characteristics and invariances, the analyzed feature extraction methods are: Haralick features, SIFT, HGM, and LETRIST. The performance of these techniques has been compared with a baseline method that computes a Histogram of Pixel Intensities.
Baseline: Histogram of Pixel Intensities. This method quantizes the pixel intensities of a grayscale image into a small number of ranges and generates a histogram descriptor by counting the occurrences in each range.
Haralick. A Gray Level Co-occurrence Matrix (GLCM) is used to count the co-occurrences of gray levels by considering the relation of each pixel with a neighbor pixel at a parametric distance. The co-occurrence is computed in 4 directions and from each generated matrix, 13 statistics are computed from the distribution of the gray-level co-occurrences. The computed values are averaged over the 4 directions to obtain a final descriptor that is robust to rotations.
SIFT. This algorithm can detect keypoints at multiple scales by using scale-space filtering produced from the convolution of Gaussian kernels at various scales. The keypoints extracted from training normal images are clustered to identify the most relevant normal keypoint descriptors which are used to build a dictionary of visual words. During inference, the detected SIFT keypoints are used to generate a *Visual Bag-of-Words* by associating each descriptor to the nearest word in the vocabulary. The resulting Bag-of-Words is the *feature vector* of the image.
LETRIST. This method computes the derivatives of an image by the convolution with a set of first and second directional Gaussian derivative filters. Then, the extremum responses are identified to extract rotation-invariant features. A set of linear and non-linear operators are applied to construct a set of transform features that are quantized into discrete texture codes and jointly encoded to build a histogram image descriptor.
Histogram of Gradient Magnitudes (HGM) [6]. This feature descriptor is based on the gradient magnitude of pixel intensities that indicate the amount of gray-level difference between pixels in the neighborhood. HGM is rotation-invariant because it computes the histogram using the gradient magnitudes, ignoring the gradient orientations.
The features extracted from these techniques are then classified with two popular ML anomaly detection methods: OC-SVM and Isolation Forest.
One-Class SVM (OC-SVM). This technique separates inliers (normal instances) from the outliers by finding a hyperplane that maximizes the boundary from the origin, i.e. all the observations with low similarity with respect to the training data. The hyperplane can be defined in kernel space (e.g. Radial Basis Function) inducing a non-linear surface in the feature space that allows the model to learn complex normal data distributions. The OC-SVM is characterized by

a ν parameter which can be used to fine-tune the trade-off between overfitting and generalization by allowing the decision boundary to consider a fraction of the training samples as outliers.

Isolation Forest. This method isolates observations by randomly selecting a feature and then splitting the data between the maximum and minimum values. Recursive partitioning can be represented by a tree structure and an ensemble of similar decision trees is used to isolate anomalies from the rest of the data. Assuming anomalies are separate from normal data, random partitioning produces shorter paths for the anomalous samples. The average path length over a forest of random trees is a measure of normality and can be used to separate normal and anomaly samples.

Regarding DL anomaly detection techniques, many architectures and models have been proposed in the literature [2]. In this work, two deep anomaly detection models have been tested: Deep SVDD, and Deep SAD. Both methods extract features from images leveraging an encoder structure based on a LeNet-style architecture composed of 4 convolutional stages. Each stage has a *Convolutional* layer, a *Batch Normalization* layer and a final *Max Pooling* layer. The output of the encoder is fed to a *Fully Connected* layer of 144 units which defines the latent representation of the input image.

Deep SVDD. This technique trains a neural network to minimize the volume of a hypersphere that encloses the latent representations of the normal training data. This forces the network to extract the common factors of variation to map all the data points towards the center of the hypersphere. During inference, points that are mapped outside the sphere are considered anomalies. Deep SVDD is characterized by two objectives: *soft-boundary* and *one-class*. The *soft-boundary* objective penalizes only the points outside the hypersphere. The hyperparameter $\nu \in (0, 1]$ controls the trade-off between the sphere volume and the boundary violations (points mapped outside the sphere). After training, the predicted anomaly score is adjusted by subtracting the radius of the hypersphere and the sign of the resulting score discriminates inliers (normal data) from outliers. The *one-class* objective penalizes the distance of all the sample representations to a central point in the latent space without defining a sphere. After training, all the predicted scores are positive. Consequently, a threshold needs to be defined to distinguish between normal and anomaly samples.

Deep SAD [3]. This method is a generalization of the unsupervised Deep SVDD method to the semi-supervised setting. This model exploits the labeled anomaly samples available during training to better define a boundary that separates normal data and anomaly instances in the latent space. The loss term of the labeled data is weighted with the hyper-parameter η which can be tuned to emphasize ($\eta > 1$) the labeled anomaly samples with respect to the unlabeled normal data. Similarly to Deep SVDD with the *one-class* objective, an anomaly threshold is used to classify a tile as normal or anomaly based on the predicted anomaly score.

4. Evaluation

For the training of unsupervised anomaly detection models only normal samples (background tiles) are used, thus all the anomaly tiles have been discarded. All the models have been fine-tuned on the validation set, and a final quantitative and qualitative evaluation has been performed on the test set. The validation and test set contain normal and anomaly tiles.

For each feature extractor and ML classifier combination, the best configuration of hyper-parameter values is selected by assessing the performance on the validation set. Specifically, the configuration that reaches the highest F1-Score on the anomaly class of the validation set is selected as the best model.

For DL models, the tested hyper-parameter configurations are ranked by the Area Under Precision-Recall Curve (AUPRC) computed on the validation set. Then, for the configuration with the highest AUPRC, the point in the Precision-Recall Curve where the model obtains 90% recall of anomaly class is selected, to reduce the probability of missing people. The corresponding threshold is used as the *anomaly threshold* for identifying anomalies.

Here is a discussion of the quantitative evaluation and hyperparameter tuning of the analyzed ML and DL techniques.

Baseline: Histogram of Pixel Intensities. The generalization capability of this baseline method is poor because the generated his-

tograms are sensitive to changes in the intensity distributions which can vary among forest types. This method relies on the presence of many nearby targets in larger tiles. This is a big weakness because, in a more realistic scenario, very few targets may appear in the frame.

Haralick. Some Haralick features are similar for anomaly and normal tiles, meaning that they are not relevant and can introduce noise. Better results are obtained when the most discriminating features are selected. However, the best subset may change based on the flight types. This method is ineffective for this anomaly detection task because the extracted global features are too generic to identify small targets.

SIFT. The implemented Bag-of-Words technique is usually adopted for supervised classification tasks where the vocabulary comprises words from all the involved classes. In an unsupervised setting, there is no word associated with the anomaly class. Consequently, the performance is poor because every keypoint detected on anomaly tiles is associated with the nearest word which, by construction, represents a cluster of normal keypoints.

LETRIST. Applying the feature selection step to the LETRIST features does not significantly influence the performance. This indicates that only a few features are relevant for identifying anomalies, but the others do not introduce much noise. The extracted features are not adequate for this anomaly detection task because, when testing on new data, the model confusion leads to the generation of a high number of False Positives.

HGM. This technique is the best among ML models. HGM relies on a simple algorithm that builds a histogram of gradient magnitudes with a range determined by the maximum and minimum values computed from each image. This approach can identify even weakly visible targets that usually have larger gradient magnitudes, especially on the borders, with respect to the gradients computed on background pixels.

Deep SVDD. Deep SVDD with both *soft-boundary* and *one-class* objectives have performance on the test set that is similar to the results obtained by the best ML techniques. This model, despite being capable of learning a better representation of normal data, may suffer from the existing knowledge of the pre-trained feature encoder.

Deep SAD. The main reason for the strong performance of Deep SAD is the ability to exploit anomaly samples during training. Anomalies are used to fit the sphere around normal data better forcing it to exclude the available anomaly examples. The results obtained with Deep SAD prove that introducing a small number of anomaly instances (e.g. 5%, or 343 samples in this case) in the training set strongly improves the anomaly detection capability.

Table 1 shows the performance of the best combination of each textural feature extractor, along with the DL models. For each method, the tile size is also reported to indicate the granularity that can be obtained when generating anomaly heatmaps. Independently from the feature extractor, OC-SVM always obtains worse results with respect to Isolation Forest. However, in some cases, the difference between the two classifiers is marginal (e.g. SIFT loses 0.4% F1-Score with OC-SVM); in other cases, the loss is remarkable (e.g. HGM loses 11% F1-Score with OC-SVM). DL models, despite relying on a simple LeNet-type architecture, on average present better performance than traditional ML methods accordingly to the F1-Score on anomaly class. The performance of ML methods mainly depends on the discriminating power of the extracted textural features. This is proved by HGM which obtains a remarkably better performance (73.5% F1-Score) with respect to the range of results obtained by other ML techniques (56.1%-65.1% F1-Score). Deep SAD demonstrates that the semi-supervised setting allows for significant improvements in anomaly detection performance despite the model being trained with few target examples.

Finally, Figure 2 shows examples of heatmaps generated by the best model, Deep SAD. The results are promising since the model can correctly highlight all the targets without confusing trees with similar textures.

5. Conclusions

Several anomaly detection techniques have been evaluated on the available data set and their performance has been extensively analyzed. For each method, the influence of the parameters on the results has been accurately studied. During the performance assessment, the character-

Model	Tile size	Test set, anomaly			Test set, background		
		Precision	Recall	F1	Precision	Recall	F1
Baseline, IFOR	216	56.8%	76.3%	65.1%	87.3%	73.8%	80.0%
Haralick, IFOR	192	69.8%	55.6%	61.9%	81.6%	89.1%	85.2%
SIFT BoW, IFOR	96	43.5%	86.4%	57.8%	88.9%	49.2%	63.4%
LETRIST, IFOR	96	43.7%	78.4%	56.1%	84.8%	54.3%	66.2%
HGM, IFOR	96	88.4%	62.9%	73.5%	85.2%	96.3%	90.4%
DSVDD one-class	96	73.2%	70.6%	71.9%	87.0%	88.3%	87.6%
DSVDD soft-bound	96	64.2%	79.1%	70.9%	89.5%	80.1%	84.5%
Deep SAD	96	**95.4%**	**90.0%**	**92.6%**	**95.6%**	**98.1%**	**96.8%**

Table 1: Evaluation results on the test set for all the models. Metrics are computed separately for the background and anomaly classes. The tile size of the best configuration is indicated for each model.

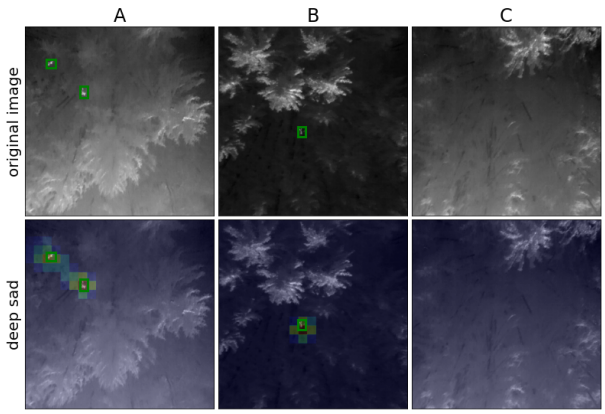

Figure 2: Anomaly heatmaps generated by Deep SAD. The model correctly identifies the targets without confusing the background. Green boxes indicate visible people.

istic of the data set have been taken into account and their influence on the methods has been discussed.
The results prove that DL models on average are more powerful than ML models (10% higher F1-Score). HGM is the only exception that obtains 73.5% F1-Score which is comparable to the performance of Deep SVDD with 71.9% F1-Score on the anomaly class of the test set. However, HGM is limited by a lower recall (62.9%) indicating that many targets are missed.
It has also been demonstrated that the availability of a few anomaly examples during training can be exploited by DL models to significantly improve anomaly detection performance. Deep SAD can obtain 92.6% F1-Score when only 5% anomaly samples are added to the training set. This result is characterized by a high recall which indicates that only a few targets might be missed during a real rescue operation.
The results are very promising and prove that approaching the problem with an anomaly detection task is suitable for detecting people from drone imagery in SAR missions. Future works will focus on applying the most promising studied techniques to images collected from real past SAR missions.

References

[1] Anne Humeau-Heurtier. Texture feature extraction methods: A survey. *Ieee Access*, 7:8975–9000, 2019.

[2] Bahram Mohammadi, Mahmood Fathy, and Mohammad Sabokrou. Image/video deep anomaly detection: A survey. 2021.

[3] Lukas Ruff, Robert A. Vandermeulen, Nico Görnitz, Alexander Binder, Emmanuel Müller, Klaus-Robert Müller, and Marius Kloft. Deep semi-supervised anomaly detection, 2020.

[4] Sasa Sambolek and Marina Ivasic-Kos. Automatic person detection in search and rescue operations using deep cnn detectors. *IEEE Access*, 9:37905–37922, 2021.

[5] David C. Schedl, Indrajit Kurmi, and Oliver Bimber. Data: Search and rescue with airborne optical sectioning. 6 2020.

[6] Monika Sharma and Hiranmay Ghosh. Histogram of gradient magnitudes: A rotation invariant texture-descriptor. pages 4614–4618. IEEE, 9 2015.