ON THE CLASSIFIER PERFORMANCE FOR SIMULATION BASED DEBRIS DETECTION IN SAR IMAGERY

S. Kuny [1,2], H. Hammer [1], K. Schulz [1]

[1] Fraunhofer IOSB, Institute of Optronics, System Technologies and Image Exploitation, Ettlingen, Germany - (silvia.kuny, horst.hammer, karsten.schulz)@iosb.fraunhofer.de
[2] Institute of Photogrammetry and Remote Sensing IPF, Karlsruhe Institute of Technology KIT, Germany

Commission I, WG I/3

KEY WORDS: SAR simulation, debris, damage detection, texture features, classifier performance.

ABSTRACT:

Urban areas struck by disasters such as earthquakes are in need of a fast damage detection assessment. A post-event SAR image often is the first available image, most likely with no matching pre-event image to perform change detection. In previous work we have introduced a debris detection algorithm for this scenario that is trained exclusively with synthetically generated training data. A classification step is employed to separate debris from similar textures such as vegetation. In order to verify the use of a random forest classifier for this context, we conduct a performance comparison with two alternative popular classifiers, a support vector machine and a convolutional neural network. With the direct comparison revealing the random forest classifier to be best suited, the effective performance on the prospect of debris detection is investigated for the post-earthquake Christchurch scene. Results show a good separation of debris from vegetation and gravel, thus reducing the false alarm rate in the damage detection operation considerably.

1. INTRODUCTION

Natural disasters, in particular earthquakes, cause a strong demand for a fast and reliable detection of structural damages. Due to the independence of weather and lighting conditions and the consequentially ensured image availability, many approaches are based on SAR imagery, occasionally in combination with ancillary data (Tao, 2016). However, the likely and rather challenging case of having neither a pre-event image nor additional data available is treated less often (Balz, 2010; Gong, 2016).

In SAR imagery, the most prominent indication for structural damages is the signature caused by heaps of debris surrounding the buildings. Due to its coarse texture, debris can be separated rather well from other signatures caused by urban formations. There are several sources, though, most importantly high vegetation and gravel, that feature a very similar texture in SAR images and thus make the debris detection approach considerably more difficult.

Previous work addressed the search for suitable textural features to describe these types of textures and the advantages of using simulated data as training samples for classification purposes. Essentially, this entails the prospect of creating generic samples, which are unaffected by random factors and independent from the actual SAR image that is to be investigated. The chosen feature set was described in (Kuny, 2015) and consists of Haralick features and some statistics of the first order. It was demonstrated by means of a multidimensional scaling that there is but some extent of overlap in the feature space regarding the signatures of debris, vegetation and gravel, and that the chosen feature set is rather capable to distinguish between the classes (Kuny, 2016a). Using a TerraSAR-X High Resolution Spotlight image of the post-earthquake Christchurch (New Zealand) scene as test data, it was shown that the major sites of debris, e.g. caused by a collapsed building, can be detected via a screening process (Kuny, 2014). Preliminary work on the classification of debris and vegetation demonstrated promising results using a random forest classifier (Kuny, 2016b).

The aim of this paper is to verify the use of a random forest classifier for the separation of the signatures of debris and vegetation in view of a simulation based training environment. For this purpose, two alternative popular classifiers, a support vector machine (SVM) and a convolutional neural network (CNN), are deployed and compared regarding their classification performance.

2. DATA

The data set used in this study consists of a High Resolution Spotlight 300MHz TerraSAR-X image, with a pixel size of 45.47 cm x 85.72 cm and an incidence angle of 47.38° (see Figure 1). It was recorded 32 hours after the February 2011 earthquake took place, destroying large parts of the inner city.

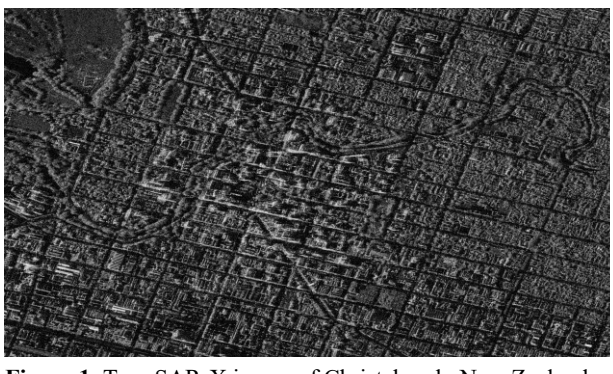


Figure 1. TerraSAR-X image of Christchurch, New Zealand.

Figure 2. Reference map regarding classes debris (red: ground-level debris, orange: higher-level debris), vegetation (green) and gravel (blue).

A reference map was generated based on an airborne orthophoto with a resolution of 10 cm (Land Information New Zealand), taking into account the classes debris, vegetation and gravel. Since the debris detection algorithm is carried out on the slant range image, the geometry needs to be taken into account in the reference map. Image registration is achieved by rasterisation of the reference map and a subsequent tie point based transformation into slant range geometry. The fact that areas of higher-level debris as well as high vegetation are projected towards near-range, is taken into account by a corresponding component shift in the case of vegetation versus a component stretch for higher-level debris. The final reference map is depicted in Figure 2.

For the evaluation process a shadow mask was generated using a LoD2.5 3d city model of the Central Business District of Christchurch and the SAR simulator CohRaS®. The model was generated by PLW Modelworks using optical imagery from 2010, thus providing pre-event conditions.

Test data

The test samples were acquired by manual extraction of 14 verified debris sites and several vegetation areas from the test area of the TerraSAR-X image, using both the reference image and visual verification to make sure there is no blockage due to neighbouring buildings. Corresponding to the training data these areas were then parcelled into 11 x 11 pixel samples, resulting in a test set of 1000 samples each for debris and vegetation. For the assessment of the classification performance, class gravel was not included for two reasons: The test area does not provide enough areas of reference signatures; secondly and more importantly, feature space has proven that debris and vegetation are much more entwined and thus classification performance is up to the separation of these two classes.

3. DEBRIS DETECTION ALGORITHM

Previously, a debris detection algorithm for a single post-event HR SAR image was developed using exclusively synthetic training samples. The general workflow of the algorithm is visualised in Figure 3, showing the consecutive processing steps. A screening step achieves the localisation of all debris-like texture in the post-event image, which - it was found - also involves the texture of vegetation and gravel. Consecutively, these classes are separated by a classification process. For a more detailed description of the algorithm we refer to (Kuny, 2016b).

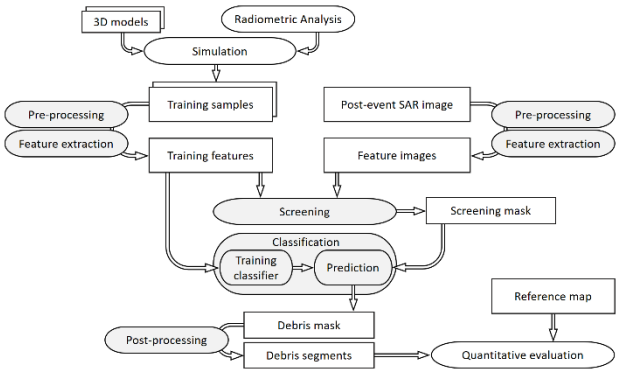


Figure 3. Workflow of the debris detection algorithm.

4. CLASSIFIER TRAINING

Depending on the classification problem, factors such as the choice of classifier can have a large impact on the performance. To validate the use of a random forest for differentiating debris from similar textures, a performance comparison between the chosen random forest classifier and two alternative classifiers (SVM and CNN) is conducted. It was refrained from including gravel as a debris-like texture, since the results on the test scene suggest vegetation to be the far more challenging factor.

The training of the classifiers is conducted using synthetic samples exclusively so as to remain conform with the damage detection approach, and thus providing for a data set that is unaffected by random factors, indefinitely expandable, and independent from the actual SAR image that is to be investigated. The process of simulating generic, radiometrically correct SAR textures employing the SAR simulation tool CohRaS® (Hammer, 2009) was already described in previous studies (Kuny, 2016a).

In order to obtain a large enough data basis for the training process, various 3d models were generated and simulated for several aspect angles. Subsequently, 11 x 11 pixel samples were extracted. Since surrounding signatures in real SAR imagery are random and thus cannot be simulated, there was made a point of using mainly sample windows located fully inside the signature, thus relying fully on the texture characteristics. Note that the use of synthetically generated training samples facilitates the establishment of a perfectly uniform class representation.

The effective training data set consists of 1000 simulated samples, 500 for each class, which is considered sufficient for the training of the random forest and the SVM. However, since the training of a CNN requires significantly more data, the set was extended by additional simulations to a total of 14.000 samples for the training of this classifier exclusively. It was found that both SVM and CNN benefit strongly from an energy normalisation of the input samples. Hence, both training and test data were normalised as specified by

$$f_{norm} = \frac{f}{\sqrt{\sum f^2}}.$$

Since there is no observable benefit for the random forest classification, though, and also to provide a comparability to the debris detection approach, the original input samples are used for the random forest classification.

In the following, specifics on the implementation and the training process regarding the three classifiers are described.

4.1 Support Vector Machine

As is to be expected the used feature set (21 features: Haralick and first order statistics) prove to be correlated to some extent. Depending on the classifier, redundant features (multicollinearity) and irrelevant features may cause overfitting and reduce the model performance, as well as lead to an unnecessarily high computational load. A random forest is robust to redundancy in the feature set; however, in the case of an SVM, the concept of feature reduction is crucial. Hence an impartial discriminant analysis was conducted. For the performance assessment a reduced set of 6 features was chosen based on the results of a sequential feature selection using Mahalanobis distances.

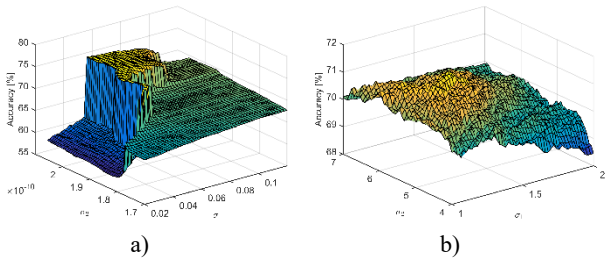
Further, SVM algorithms are not scale invariant as are for example tree based algorithms, which leads to an unbalanced feature influence in cases that individual features do not have a similar range of values. For that reason, the features are standardised before they are fed to the SVM, which involves a scaling to have zero-mean and unit-variance. Also, an energy normalisation of the feature vectors is conducted.

For the performance comparison a Radial Basis Function (RBF) kernel SVM is employed, which consists of a two-step SVM developed at Fraunhofer IOSB. Firstly, 2-class SVMs are used to discriminate all pairs of classes (a pre-classification) and secondly 1-class SVMs determine the class memberships based on the resulting new feature vector. It is a rather universal tool, where much of its power consists of the aptitude to handle more than two classes, which for the problem at hand is non-relevant. For a detailed description of the method see (Middelmann, 2006).

The RBF kernel, which is defined as

$$k(x, y) = e^{-\frac{\|x-y\|^2}{\sigma^2}},$$

where x and y denote sample data (support vectors) and σ is the standard deviation, maps the sample data to a high dimensional space. By maximising the minimal distance between the supporting vectors and the separating hyper plane the ideal kernel parameter is identified. The SVM mainly uses three hyper-parameters: the kernel parameter σ_2 of the 2-class SVMs, the kernel parameter σ_1 of the 1-class SVMs and a reject threshold. To identify the ideal values for the hyper-parameters σ_1 and σ_2 a grid search is conducted aiming to locate the global maximum. Figure 4 a) visualises the grid search including the located maximum. For comparison, Figure 4 b) demonstrates the grid search using Fast Fourier Transform-based features instead of the selected texture features. However, the maximal reached accuracy cannot compete.



a) b)

Figure 4. Grid search of the RBF kernel SVM with a) using the defined feature set and b) using Fast Fourier Transform-based features.

4.2 CNN

For many fields of application CNNs prove to be the most powerful tool available, and hence need to be considered for the task at hand. A notable difference to the described classification approach with random forest or SVM is the input data. Whereas random forest and SVM classification are based on the introduced set of extracted texture features, the CNN, as a feature extractor of its own, is fed with the image samples.

The nature of the problem under consideration suggests a rather shallow architecture model, thus focussing on models with no more than two convolution layers. Deeper structures were tested, however, the performance was bad, also due to the involved substantial overfitting of the model. Pooling can be a means to reduce overfitting. However, in this case an abundant use of pooling layers is not an option, since the 11 x 11 pixel sample size is very small to begin with and further sub-sampling would result in a significant loss of information. Other measures against overfitting include the use of dropout layers or simplifying the model. Both methods were explored, with the conclusion that a shallow structure without dropout layers leads to a good training and simultaneously prevents overfitting.

The best results were achieved with an architecture as follows. Two 2-d convolutional layers are employed using 10/20 filters of size 3 x 3 and a stride (step size for roaming the input) of 1. Also included is a zero padding, which implies the padding of the borders to enforce a preservation of the input size. Further, one maxpooling layer with a 2 x 2 pooling region and a stride of 2 was installed. It operates by breaking down the input into rectangular sectors and returning each maximal value. The architecture concludes with two fully-connected layers and the application of a softmax-function to the output. Finally, the classification layer computes the cross entropy loss.

Regarding the process of training, 95% of the 14.000 simulated samples were employed as training data whereas the remaining 5% were used for validation purposes. An initial learning rate of 0.1 with a gradual decay every 25 epochs proved to be suitable. The training iterations were conducted using a number of 256 mini-batches and was continued until the mini-batch loss dropped to a value of 0.0001. The development of the mini-batch accuracies throughout the training can be observed in Figure 5.

4.3 Random Forest

For reasons of comparability, the proceedings and settings regarding the training of the random forest classifier correspond to those described in previous works (Kuny, 2016b). Out-of-bag Error (OOB) estimates are employed as measure of prediction error, thus avoiding the need for an independent validation dataset. Tuning showed a minimum OOB error for a number of 17 features.

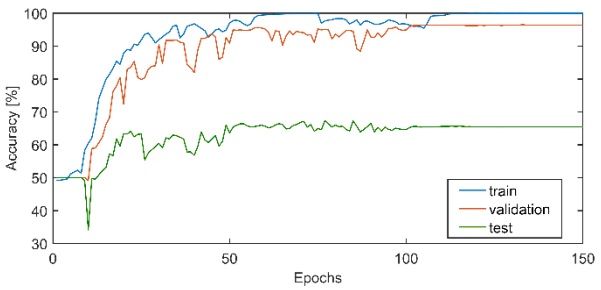


Figure 5. CNN training accuracies per epochs.

For the classification fully grown trees are used since the computational load is manageable, however, it was found that reducing tree depth moderately does not decrease the classification performance. It was also revealed that a number larger than approximately 60 trees does not improve the model performance.

5. RESULTS

5.1 Classifier Performance

Since the introduced test classes are perfectly balanced, accuracy (ACC) is a valid measure in this case and hence can be used as evaluation criterion. Table 1 shows the classification results for classes debris (C_D) and vegetation (C_V) attained by the three different classifiers, whereas corresponding performance measures are listed in Table 2.

		Random Forest		SVM		CNN	
		C_D	C_V	C_D	C_V	C_D	C_V
Test	C_{TD}	967	33	935	65	483	517
	C_{TV}	432	568	414	586	130	870

Table 1. Confusion matrices regarding classification results of test data.

	ACC [%]	TPR [%]	PPV [%]
Random Forest	76.8	96.7	69.1
SVM	76.1	93.5	69.3
CNN	67.7	48.3	78.8

Table 2. Classifier performance.

The classification results show a good performance for both random forest and SVM, with 76.8% and 76.1% ACC respectively. Considering the limits of the selected set of texture features with regard to a separability of the two classes (Kuny, 2016a) this is a satisfactory result. Since both classifiers were fed with a feature set of the same information content, the similar results seem conclusive.

The CNN approach achieved an ACC of 67.7%, which is rather poor in comparison. Considering the impressive performance of CNNs in other fields of application this result initially is quite unexpected. However, the power of a CNN stems from learning the entirety of a target, including its form and borders/surroundings. Bearing this in mind, the action of limiting the input to 11 x 11 pixel samples containing exclusively debris texture, limits the CNNs feature extraction capacity severely. This is assumed to be the main reason for the rather poor results of the CNN approach. In summary, these results can verify the aptitude of a random forest classifier in the context of this damage detection approach.

Note that this performance analysis is based on the pixel-wise classification, hence the resulting classification rates are rather conservative. It stands to reason that a component-wise classification would result in distinctly better classification rates. However, since the main aspect here is to establish a comparison to alternative classifiers, a pixel-wise classification was considered solid.

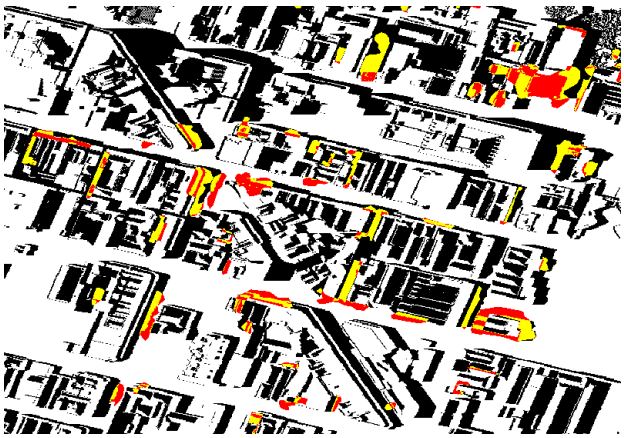


Figure 6. Assessment of shadowing rate: shadow mask (black) overlayed with ground-level reference debris (red) in slant range geometry (yellow denotes overlapping areas).

5.2 Damage detection performance

The typically flat incidence angle ($\theta = 47.38°$) of the TerraSAR-X image acquisition leads to a considerable amount of shadowing, particularly in areas of high building development such as in the test scene. As a consequence, there are many debris occurrences that are located partly or entirely in the shadow of a building. The reference map, however, as independent data set, does not provide this information. To approximate the extent of shadowing in the test scene, or rather to estimate the rate of debris occurrences that are not in line of sight of the sensor, the shadow mask of the inner city of Christchurch is introduced. As a matter of course, this approach neglects the case of reduced shadowing due to the collapse of buildings. However, quantitatively, these incidences are considered scarce enough to be disregarded here.

Since the 3d city model covers the test scene only partially, a representative cut-out of the scene (approximately 0.3 km^2) was defined for an assessment of the shadowing rate regarding ground-level debris occurrences. The shadow mask for this area reveals that for the acquisition geometry in question, the shadow coverage amounts to a total of 38.0 % of the cut-out scene. Figure 6 shows the shadow mask (black) of this cut-out of the test scene overlayed with the reference mask of the ground-level debris (red), thus marking areas (yellow) that refer to debris occurrences located in the shadow areas. According to this, a total of 54.7% of the ground-level reference debris occurrences are not in line of sight of the sensor, and consequently cannot be detected using this acquisition geometry. For the quantitative evaluation, the consequences are bound to be significant. It is to be expected that an approximate of 55% of the reference ground-level debris components cannot be detected in the screening process.

With the confirmation that the random forest is a suitable choice, the damage detection algorithm is assessed on the Christchurch test area. The screening step, isolating debris-like texture, narrows down the working area to less than 7% of the test area. The random forest classification, applied on the screening mask, yields the predicted classes debris, vegetation and gravel. Figure 7 depicts the final component-wise classes after post-processing operations with component-wise majority voting. The predicted class debris contains a total of 822 components with class vegetation and class gravel comprising 1059 and 19 components, respectively.

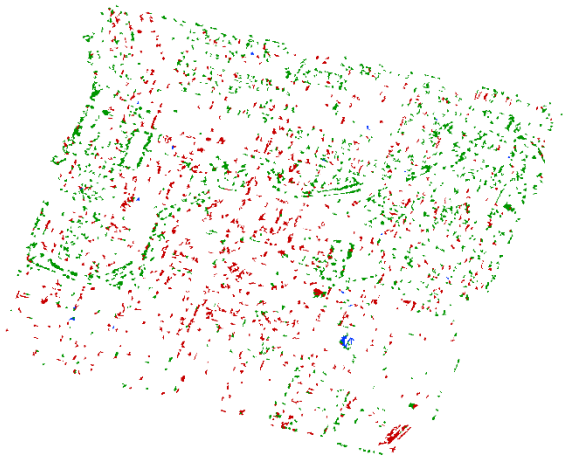


Figure 7. Christchurch test area: predicted classes debris (red), vegetation (green) and gravel (blue) for a closed world scenario.

The following provides a detail analysis of exemplary points of interest: The first example contains a large debris site pictured in Figure 8 a). The site is in full line of sight of the SAR sensor and is the result of a fully collapsed building. According to the reference map, the site covers almost 2000 m^2 (3474 pixels) with almost no remaining vertical elements at the scene. Examining the screening and classification results versus the reference map in Figure c), d), and e) several observations can be made. The algorithm results for this area show that the major part of this large debris site was found in the screening process, also catching some of the surrounding smaller debris sites. Apart from these areas, the screening result contains also several areas that belong to non-debris sources, most prominently the vegetation area in the upper right corner. The predicted classes show that these areas of high vegetation are distinguished rather well by the trained classifier, even separating single trees that are located adjacent to debris. While a certain loss regarding actual debris can be observed, most of the large debris site is classified correctly, as well as the elongated site on the opposite side of the road, representing heaps of debris in front of a still standing building.

The second example specifies a rectangular park area framed by high trees and surrounded by residential housing (see Figure 9).

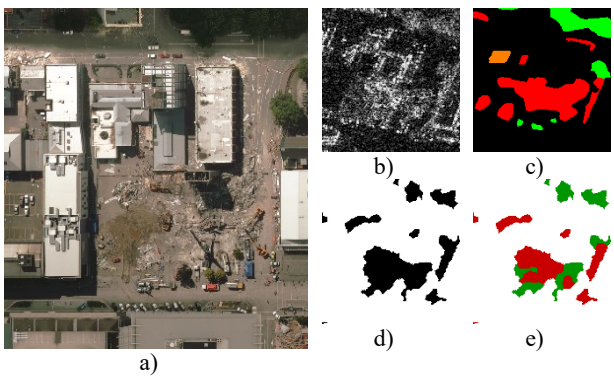


Figure 8. Exemplary debris site: a) Optical image, b) SAR image, c) reference map, d) screening result, and e) classification result.

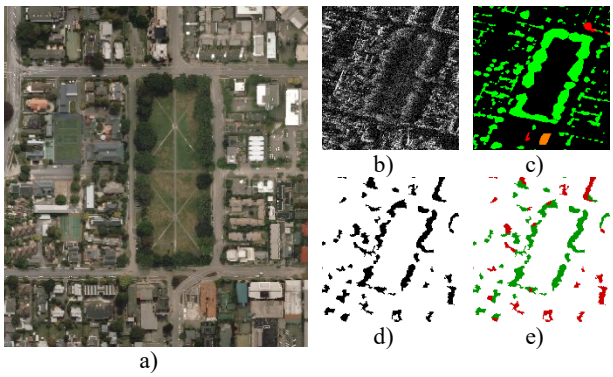


Figure 9. Exemplary area of vegetation: a) Optical image, b) SAR image, c) reference map, d) screening result, and e) classification result.

It can be observed that large enough vegetation, such as the park area, is almost entirely included in the screening mask, but can be separated effectively by the classifier. Further, this area can demonstrate the two main difficulties involving residential areas. As is common for residential areas, there is much vegetation that is rather small but wide-spread. Secondly, there are many small-scale constructions, such as balconies and backyard structures, which for the resolution at hand can lead to a texture similar to that of debris. These incidences located in close proximity result in screening components that are large enough not to be disregarded in the filtering step. Consequently, this results in a large amount of screening components, which due to the closed world assumption without reject option cannot be classified correctly.

For a direct assessment of the random forest performance, a confusion matrix is analysed, which is based on the components of the screening matrix. Table 3 shows this confusion matrix regarding the true classes debris (C_{TD}), vegetation (C_{TV}), gravel (C_{TG}) and other (C_{TO}), where other denotes areas that are unspecified in the reference map, thus appertaining to signatures of unknown source. Since the true classes of the screening map areas are highly imbalanced, with 150 debris instances, 1254 vegetation instances and only 7 gravel instances ACC is not an ideal evaluation measure and as such is not included in the table. A far more useful depiction of the confusion matrix is given by the true positive rate (TPR), the positive predictive value (PPV) and the F1 score. The TPR of the classes debris and vegetation shows (with 72.7 % and 71.1 %) a rather high percentage of correctly classified instances, whereas the PPV values reflect the fact that a significant number of vegetation instances was falsely classified as debris. The rather low number of 13.3% PPV for class debris, however, is condoned in order to have a high TPR value in return since the cost of misclassified debris instances is ranked much higher than cost of instances falsely classified as debris. Though the rates for class gravel seem rather low and are debatable due to the very low number of gravel instances, it is to be noted that misclassification cases seem to predominate between the classes gravel and vegetation, whereas the separation from debris is successful.

		Predicted Class				
		C_D	C_V	C_G	TPR	PPV
True	C_{TD}	109	41	0	72.7%	13.3%
	C_{TV}	356	899	10	71.1%	84.9%
	C_{TG}	0	4	2	33.3%	10.5%
	C_{TO}	357	115	7		

Table 3. Classification performance: confusion matrix.

A screening detection rate is computed regarding the set of 282 debris components of the reference map (see Figure 2). Effectively, the screening mask includes 128 of these components, which amounts to 45.5%. Considering the estimated 55% shadowing rate of ground-level debris occurrences, this is a fairly satisfactory rate. Hence it is warranted, that the screening mask provides a good coverage of debris occurrences in line-of-sight of the sensor. The classification of these 128 components led to 85.2% being classified correctly as debris, which corresponds to 37.7% of the entire set of reference debris components. Figure 10 shows the location of these detected components (black) versus the components that were not detected (grey), whereas Table 4 summarises the reference components and their detection rates regarding the process of screening and classification.

The quantitative results demonstrate a good performance of separating vegetation and gravel from the signature of debris, thus reducing the false alarms in the damage detection operation markedly.

	Components		
	detected	**undetected**	rate [%]
Screening	128	154	45.5
Classification	109	173	37.7

Table 4. Detection rate of reference debris components.

Figure 10. Reference debris components subdivided into those that were detected (black) and those that were not (grey).

6. CONCLUSION

The aim of this paper was to verify the use of a random forest classifier for the separation of the signatures of debris and vegetation. For this purpose, two alternative popular classifiers, a support vector machine (SVM) and a convolutional neural network (CNN), were deployed and compared regarding their classification performance. In the course of this, classifier specific requirements were taken into account. Results show that the random forest, though the most straightforward of the classifiers, performs better than either SVM or CNN in this specific case. Whereas the SVM, fed with the same set of statistical texture features as the random forest classifier, reaches similar classification accuracies (about 76%), the CNN results show distinctly lower rates (67%).

With the conclusion that the random forest classifier is the most suitable, this paper also presented the most recent results on the prospect of debris detection for the post-earthquake Christchurch scene. This includes a quantitative evaluation on the basis of reference data that is derived from an RGB orthophoto and as such does not represent the shadowing conditions of the post-event SAR image. In this context, a simulated shadow mask of the scene was employed since it enables an assessment of the extent of shadowing present in the test scene. We also provided detail analyses of some exemplary points of interest. The quantitative results demonstrate a good performance of separating vegetation and gravel from the signature of debris, thus reducing the false alarms in the damage detection operation markedly.

REFERENCES

Balz, T., and Liao, M., 2010: Building damage detection using post-seismic high-resolution SAR satellite data. *International Journal of Remote Sensing 31*, 3369–3391.

Gong, L., Wang, C., Wu, F., Zhang, J., Zhang, H., and Li, Q., 2016: Earthquake-induced building damage detection with post-event sub-meter VHR TerraSAR-X staring spotlight imagery. *Remote Sensing 8*, 11, 887.

Hammer, H., Schulz, K., 2009: Coherent Simulation of SAR Images, *Proceedings of SPIE, Image and Signal Processing for Remote Sensing* XV, doi: 10.1117/12.830380, Vol. 7477, pp. 74771G-1-9.

Kuny, S., Schulz, K., 2014: Debris detection in SAR imagery using statistics of simulated texture. *8th IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS)*, vol.4, 1-4.

Kuny, S., Hammer, H., Schulz, K., 2015: Discriminating between the SAR signatures of debris and high vegetation. *International Geoscience and Remote Sensing Symposium (IGARSS),* 473-476.

Kuny, S., Hammer, H., Schulz, K., 2016a: Assessing the suitability of simulated SAR signatures of debris for the usage in damage detection. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences (ISPRS),* Volume XLI-B3, 877-881.

Kuny, S., Hammer, H., Schulz, K., Hinz: S., 2016b: Towards a reliable detection of debris in high resolution SAR images of urban areas. *11th European Conference on Synthetic Aperture Radar (EUSAR)*, 1015-1018.

Land Information New Zealand: https://data.linz.govt.nz/layers/, last visited: 2020-03-15.

Middelmann, W., Ebert, A., and Thönnessen, U., 2006: Assessment of a novel decision and reject method for multi-class problems in a target classification framework for SAR scenarios. *Proceedings of SPIE, Algorithms for Synthetic Aperture Radar Imagery XIII 6237,* 200–208.

Tao, J., and Auer, S., 2016: Simulation-based building change detection from multiangle SAR images and digital surface models. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 9*, 3777–3791.