A Deep Convolutional Neural Network for Semantic Pixel-Wise Segmentation of Road and Pavement Surface Cracks

Mark David Jenkins*, Thomas Arthur Carr*, Maria Insa Iglesias*, Tom Buggy* and Gordon Morison*
*School of Engineering and Built Environment
Glasgow Caledonian University
Glasgow, United Kingdom
Email: gordon.morison@gcu.ac.uk

***Abstract*—Deterioration of road and pavement surface conditions is an issue which directly affects the majority of the world today. The complex structure and textural similarities of surface cracks, as well as noise and image illumination variation makes automated detection a challenging task. In this paper, we propose a deep fully convolutional neural network to perform pixel-wise classification of surface cracks on road and pavement images. The network consists of an encoder layer which reduces the input image to a bank of lower level feature maps. This is followed by a corresponding decoder layer which maps the encoded features back to the resolution of the input data using the indices of the encoder pooling layers to perform efficient up-sampling. The network is finished with a classification layer to label individual pixels. Training time is minimal due to the small amount of training/validation data (80 training images and 20 validation images). This is important due to the lack of applicable public data available. Despite this lack of data, we are able to perform image segmentation (pixel-level classification) on a number of publicly available road crack datasets. The network was tested extensively and the results obtained indicate performance in direct competition with that of the current state-of-the-art methods.**

I. Introduction

Structural monitoring has traditionally been a highly manual task carried out by specially trained engineers or technicians. The rise of deep learning and computer vision technology has started to influence this field heavily and has lead to a dramatic increase in demand for automation or assistive technologies in structural monitoring. Within the field of structural monitoring, the analysis of roads and pavements, specifically identification of surface cracks, is an area which can benefit hugely from the application of deep learning and computer vision. While automated bounding-box style detection of such cracks in images is useful, the real value lies in accurate analysis of the precise shape, size and orientation which would require semantic segmentation to be achieved accurately.

The increase in the power and availability of GPU technology has opened the door for Deep Learning to be applied to a variety of tasks in both signal [1], [2] and image processing [3], [4]. One such task is that of semantic segmentation of images which is a vital component of a number of computer vision and image processing tasks. Applications of semantic segmentation range from driverless vehicles [5], [6] to medical imaging [7], [8] and the demand for highly accurate algorithms is rapidly increasing.Segmentation is the process of assigning each pixel in an image with a predefined class label. The output of a segmentation algorithm with only two classes for example would be a binary image, typically with the same width and height as the input image, where pixels with value 0 belong to one class and pixels with value 1 belong to the other. This output mask can then be used to carry out further analysis of the image in question or could be used to augment the input image depending on the application.

Recent advances in semantic segmentation algorithms are primarily based on Convolutional Neural-Network (CNN) methods such as the popular SegNet architecture [9] which is a common encoder-decoder network used for image segmentation. A major issue with the typical CNN approach is that it tends to require a significant volume of training data to achieve high quality results. This can be problematic in several applications which would benefit from the high performance of a segmentation CNN but are limited in the volume of data which is available for training. Another key problem with the encoder-decoder framework is that the dimensionality reduction associated with the encoder (down-sampling) layers can result in loss of fine image details which may be vital for classification and cannot be recovered by the decoder (up-sampling) layers.

The development of architectures such as U-Net [10] attempt to solve this issue by allowing for much smaller training sets and thus allow CNNs to be applied to tasks which would have previously been outwith their scope. This architecture also utilises the output of each encoder layer in the corresponding decoder layer which aids in preservation of fine image detail. This makes the U-Net architecture well suited to the pavement and road crack segmentation task.

Despite the increased use of Deep Learning in computer vision tasks, there is a lack of recent work within the field of structural monitoring, specifically crack analysis, that takes advantage of these techniques [11], [12] and the majority of crack analysis work utilises techniques such as edge detection and thresholding [13], [14]. This paper presents an algorithm

ISBN 978-90-827970-1-5 © EURASIP 2018

for semantic segmentation of road and pavement surface cracks using a Convolutional Neural Network, namely U-Net. The algorithm is trained, validated and tested on the publicly available CrackForest [15], [16] dataset which consists of 118 images of surface cracks on pavement and road surfaces, taken with a hand-held camera. To ensure fair performance evaluation, the metrics used in evaluation of the CrackForest algorithm are also used in this work.

II. Network Architecture

As discussed above, the U-Net architecture [10] is designed for image segmentation tasks for which training data is limited and where significant reduction in resolution through down-sampling is undesirable. U-Net was originally designed for segmentation of biomedical images such as segmenting microscope images of cells and bacteria and segmentation of veins and capillaries in retinal images. This biomedical task and the road crack segmentation task, while originating from very different fields of study, have several very strong correlations. Both tasks are heavily limited in the volume of training data available, both tasks contain fine detail image elements which can be lost in down-sampling and both require a highly accurate segmentation output for effective analysis. Due to the unintuitive similarities in input data and output requirements between biomedical image segmentation and road crack segmentation, the U-Net architecture is ideally suited to this task.

The network architecture can be intuitively separated into two main components; the encoder section and the decoder section. The encoder section of the network operates in the typical manner commonly associated with CNNs [9]. The input image, a single channel grey-scale image in this particular application, is passed through a succession of convolutions, Rectified Linear Unit (ReLU) activation functions [17], and Max-Pooling operations. One layer of this section of the network, which will be referred to as an encoder block, operates as depicted in Figure 2a.

Given an input of shape $[W \times H \times D]$, where W, H and D are the width, height and depth of a given input, each of the components of and encoder block produces an output of shape $[(W-4)/2 \times (H-4)/2 \times (D \times 2)]$, with the exception of the first layer which maps the $D = 1$ input image to a $D = N$ feature vector where D is the minimum feature depth utilised in the network. This is more clearly illustrated in the first half of Figure 1 which shows how the encoder blocks operate on an input image of shape $[572 \times 572 \times 1]$.

For the task of whole image classification, a Softmax layer could be implemented at this stage to map the feature vector to a probability distribution with a number of elements equal to the number of classes. For the segmentation task, however, a decoder section must be added to the network to upscale the feature vectors and allow a probability distribution to be generated for each pixel in the input image rather than for the image as a whole. Typically, a block in the decoder section of a segmentation network is constructed as shown in Figure 2b.

In a standard network, the decoder layer is only influenced by the encoder network in that the shape of the convolution and up-sampling operations must be sufficient to return a feature vector of equivalent width and height as in input image, minus the border which results from the lack of padding in the convolution and max-pooling operations. The dimensions of this output feature vector will be referred to as $[W^o \times H^o \times N]$. This means that any fine image details which are not accurately captured by the lowest dimensional feature vector are permanently lost and cannot be recovered in the up-sampling stages of the decoder layer.

The U-Net architecture attempts to tackle this by allowing each decoder block access to the input feature vector of the associated encoder block. The appropriately cropped output of the first encoder block is concatenated to the feature vector in the final block of the decoder before the convolutions are carried out. This technique reintroduces any finer image details which may have been lost during the lower level encoding layers and allows the decoder layers to reconstruct these details more effectively.

The final layer of the network is a Softmax layer which takes the final $[W^o \times H^o \times N]$ feature vector and converts it to a $[W^o \times H^o \times C]$, where C is the number of segmentation classes. Essentially, this Softmax returns a probability distribution of length C for each pixel in the subsection of the input image defined by the dimensions $[W^o \times H^o]$. The final $[W^o \times H^o \times 1]$ segmentation mask can be computed by assigning each pixel a value equivalent to the index of the maximum value in its corresponding probability distribution along the C axis of the Softmax output i.e. an argmax function.

Figure 1 gives a visual representation of the U-Net architecture as defined throughout this section.

III. Proposed Configuration

Despite the suitability of the U-Net architecture to the road crack segmentation task, there are some differences between the input data that can degrade the performance. Consider a binary segmentation problem (two distinct classes) which is standard in both the biomedical field and road crack segmentation. Generally speaking, class label 0 would be background and class 1 would be the target object. In a typical microscope image of bacteria, cells or retina capillary image a large percentage of the image contains the target object which is beneficial to training. Road crack images differ from this in that the vast majority (often upwards of 95%) of pixels in an image are background, or class 0, pixels. This greatly limits the volume of useful training data available.

To overcome this challenge, the proposed algorithm trains on randomly extracted patches which are extracted from the training images pre-training. This random sampling is constrained so that a minimum of 60% of the patches contain the target class. This ratio was found to produce the best precision while minimising the false positive rate. It was noticed that the false positive rate increased with a higher percentage of patches which contain the target class. This is most likely due

ISBN 978-90-827970-1-5 © EURASIP 2018

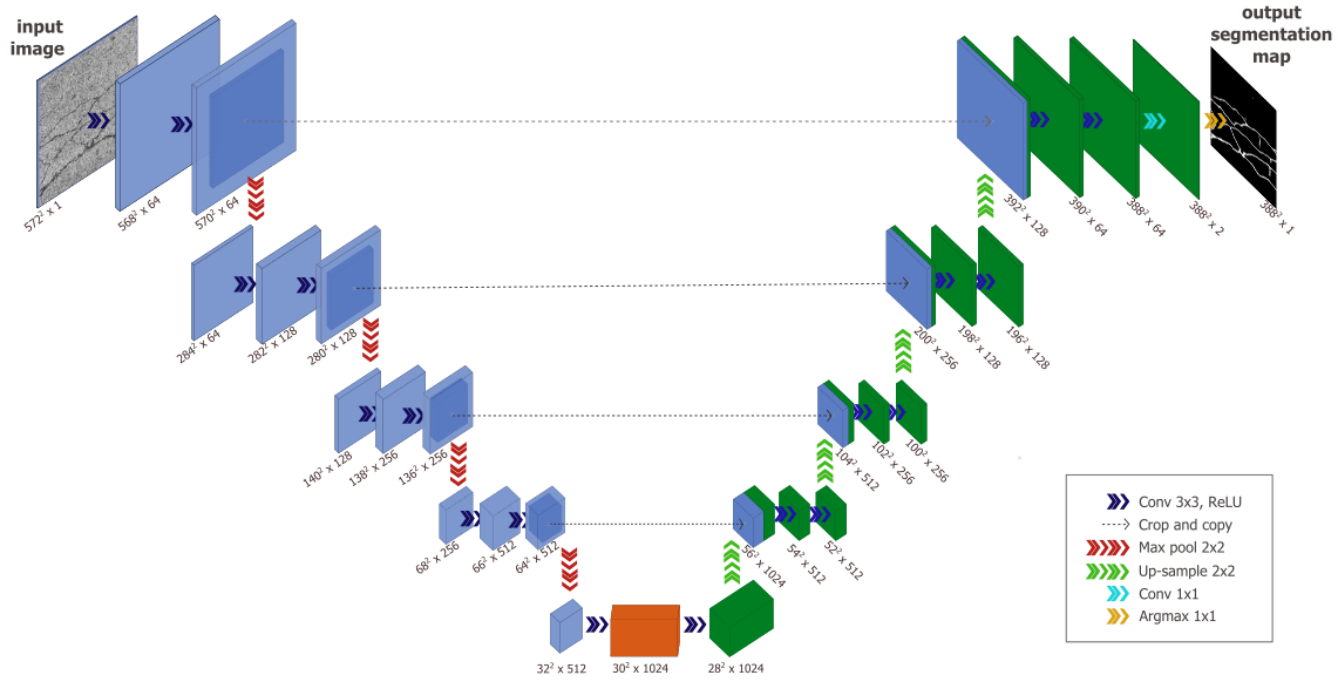
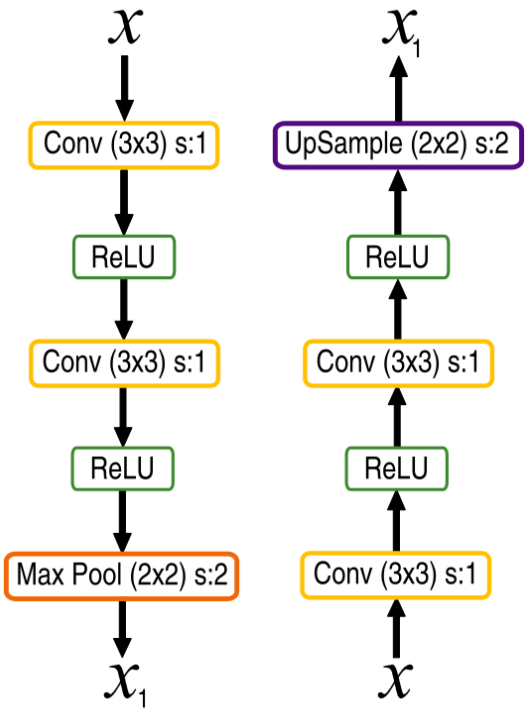


Fig. 1. Diagrammatic representation of the U-Net architecture. The Blue tiles on the left hand side represent the Encoder (down-sampling) section of the network and the Green tiles on the right show the Decoder (up-sampling) section.



(a) Encoder (Down-sampling) (b) Decoder (Up-sampling)

Fig. 2. Standard configuration of the Encoder and Decoder layers of the U-Net Architecture ($N \times N$) indicates the kernel size and s:α is the stride

to the training set lacking examples of feature dense image regions which do not belong to the target class.

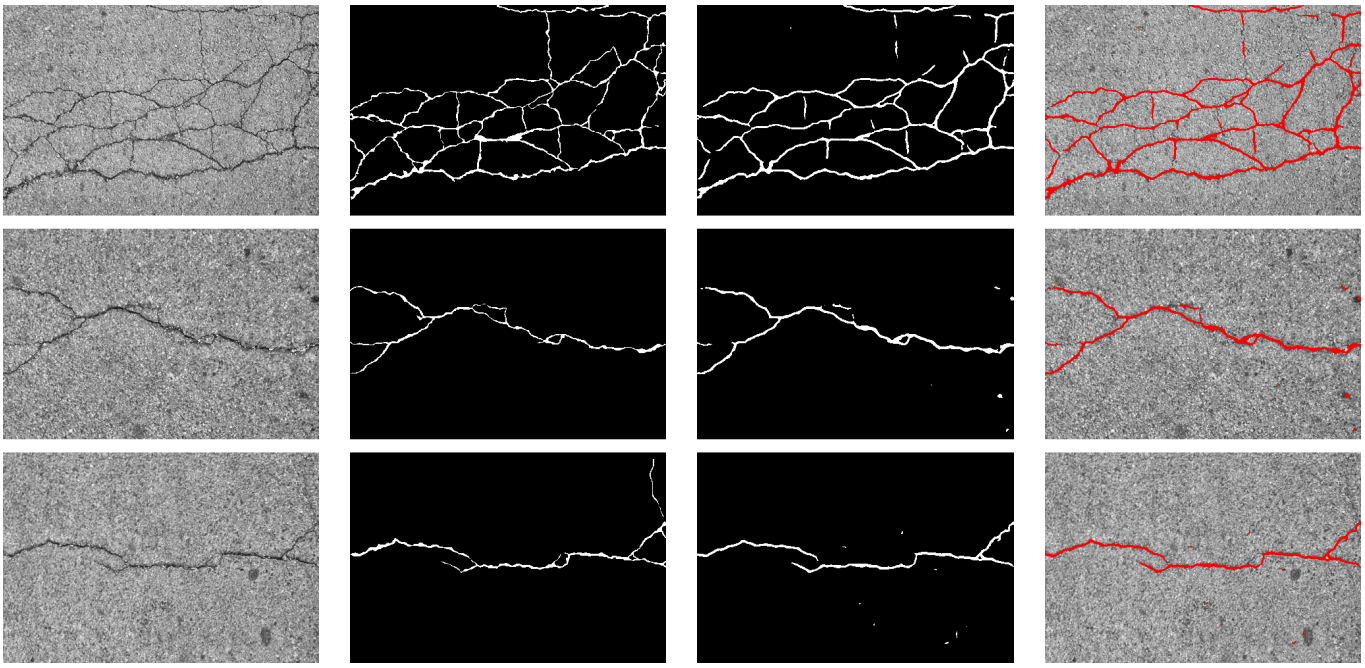
This patch based approach also has benefits in inference. Test images are processed using the sliding window method with a stride of 1. This has two distinct advantages. The first is that with minimal image padding, inference can produce a segmentation output of equal size to the input image. The second is that because inference is carried out with overlapping patches, the results can be averaged across patches which adds extra invariance to the position of target class information in the image patch.

For this particular application image patches of size [48×48] is utilised and a total of 2000 random patches are extracted from each of the training images.

IV. Evaluation & Results

The algorithm proposed here was evaluated on the Crack-Forest Dataset made available by [15], [16]. This dataset consists of 118 single channel grey-scale images of size [480×320]. Each of these images is supplied with a corresponding binary ground truth image of the same size highlighting the cracks in the image with the label value of 1. The network is trained on 100 of these images split into 80 training and 20 validation images. The remaining 18 images are used exclusively for testing. To allow for fair evaluation of the proposed method, the same three performance metrics utilised in the CrackForest work are also used in this work. These three metrics are Precision (Equation 1), recall (Equation 2) and F1-Score (Equation 3) defined as:

ISBN 978-90-827970-1-5 © EURASIP 2018



(a) (b) (c) (d)

Fig. 3. These images show; (a) three examples of pavement surface crack images from the CrackForest dataset, (b) their respective Ground Truth labels, (c) the segmentation predictions generated by the algorithm presented in this paper and (d) the overlay of these predictions on the original images for clearer evaluation.

$$Pr = \frac{TP}{TP + FP} \tag{1}$$

$$Re = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2 \times Pr \times Re}{Pr + Re} \tag{3}$$

where TP is the number of true positives, FP is the number of false positives and FN is the number of false negatives. As in CrackForest, a true positive is defined as any labelled pixel in the output segmentation mask that is within 5 pixels of a true label in the ground truth.

Table I shows the results of the proposed algorithm in comparison to another 6 crack detection algorithms. These algorithms include the baseline Canny Edge Detection method [18] as well as the following methods which are considered to be state of the art in road crack analysis: CrackTree [19], CrackIT [20], FFA [21] and two variants of the CrackForest algorithm using KNN [16] and SVM [16].

The proposed method outperforms all other methods on two of the three metrics (Precision and F1-Score) and achieves the second best result on the third metric (Recall). The algorithm outperforms the next best state of the art method (CrackForest) by a considerable 10% on the Precision metric.

The algorithm was trained, validated and tested on an NVIDIA Titan Xp GPU with 12GB of RAM. The implementation was carried out using Keras and Tensorflow and training was carried out over 100 epochs with a batch size of 34. Training takes approximately 3 hours and inference on the pre-processed test set of 18 images takes approximately 3 seconds.

Method	Precision	Recall	F1-Score
Canny [18]	12.23%	21.15%	15.76%
CrackTree [19]	73.22%	76.45%	70.80%
CrackIT [20]	67.23%	76.69%	71.64%
FFA [21]	78.56%	68.43%	73.15%
CrackForest (KNN) [16]	80.77%	78.15%	79.44%
CrackForest (SVM) [16]	*82.28%*	**89.44%**	*85.71%*
Proposed (U-Net)	**92.46%**	*82.82%*	**87.38%**

TABLE I
PRECISION (PR), RECALL (RE) AND F1-SCORE (F1) FOR A RANGE OF ALGORITHMS ON THE CRACKFOREST DATASET. THE BEST SCORE IN EACH COLUMN IS PRESENTED IN **BOLD** WHILE THE SECOND BEST IS IN *Italics*.

V. CONCLUSION

The automation of condition monitoring tasks is a field which is becoming much more prominent in the Computer Vision community. Segmentation of pavement and road surface cracks is just one component of condition monitoring which is benefiting greatly from the increase in power and availability of GPU technology. The work presented in this paper focuses on the automated semantic segmentation of surface cracks on road and pavement images. The approach taken utilises an Encoder-Decoder CNN framework known as U-Net in combination with a patch based training methodology and achieves competitive results in comparison to current state

ISBN 978-90-827970-1-5 © EURASIP 2018

of the art methods, outperforming the state of the art on two of the three evaluation metrics used.

During the evaluation of the proposed algorithm it was noted that the accuracy of the semantic segmentation was lower in instances when the target crack ran vertically through the image than in the case of horizontal cracks. It would appear from analysis of the training dataset that this is due to a significant under-representation of vertical crack images. The majority of the dataset has the main crack running horizontally and minor cracks which run vertically off of the main crack. In the future, the patch based training method proposed here will be extended to include augmentations of the input data in an attempt to equally represent cracks running at multiple angles. This should further increase the robustness of the algorithm and is expected to have a significant impact on the results.

REFERENCES

[1] N. Passalis, A. Tsantekidis, A. Tefas, J. Kanniainen, M. Gabbouj, and A. Iosifidis, "Time-series classification using neural bag-of-features," in *Signal Processing Conference (EUSIPCO), 2017 25th European*. IEEE, 2017, pp. 301–305.

[2] H. K. Vydana and A. K. Vuppala, "Residual neural networks for speech recognition," in *Signal Processing Conference (EUSIPCO), 2017 25th European*. IEEE, 2017, pp. 543–547.

[3] E. Ribeiro, A. Uhl, F. Alonso-Fernandez, and R. A. Farrugia, "Exploring deep learning image super-resolution for iris recognition," in *Signal Processing Conference (EUSIPCO), 2017 25th European*. IEEE, 2017, pp. 2176–2180.

[4] M. Tschannen, L. Cavigelli, F. Mentzer, T. Wiatowski, and L. Benini, "Deep structured features for semantic segmentation," in *Signal Processing Conference (EUSIPCO), 2017 25th European*. IEEE, 2017, pp. 61–65.

[5] M. Siam, S. Elkerdawy, M. Jagersand, and S. Yogamani, "Deep semantic segmentation for automated driving: Taxonomy, roadmap and challenges," *arXiv preprint arXiv:1707.02432*, 2017.

[6] M. Treml, J. Arjona-Medina, T. Unterthiner, R. Durgesh, F. Friedmann, P. Schuberth, A. Mayr, M. Heusel, M. Hofmarcher, M. Widrich *et al.*, "Speeding up semantic segmentation for autonomous driving," in *ML-ITS, NIPS Workshop*, 2016.

[7] F. Jiang, A. Grigorev, S. Rho, Z. Tian, Y. Fu, W. Jifara, K. Adil, and S. Liu, "Medical image semantic segmentation based on deep learning," *Neural Computing and Applications*, pp. 1–9.

[8] B. Kayalibay, G. Jensen, and P. van der Smagt, "Cnn-based segmentation of medical imaging data," *arXiv preprint arXiv:1701.03056*, 2017.

[9] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[11] H. Cheng, J. Wang, Y. Hu, C. Glazier, X. Shi, and X. Chen, "Novel approach to pavement cracking detection based on neural network," *Transportation Research Record: Journal of the Transportation Research Board*, no. 1764, pp. 119–127, 2001.

[12] B. J. Lee, H. Lee *et al.*, "Position-invariant neural network for digital pavement crack analysis," *Computer-Aided Civil and Infrastructure Engineering*, vol. 19, no. 2, pp. 105–118, 2004.

[13] Y.-C. Tsai, V. Kaul, and R. M. Mersereau, "Critical assessment of pavement distress segmentation methods," *Journal of transportation engineering*, vol. 136, no. 1, pp. 11–19, 2009.

[14] S. Chambon and J.-M. Moliard, "Automatic road pavement assessment with image processing: review and comparison," *International Journal of Geophysics*, vol. 2011, 2011.

[15] L. Cui, Z. Qi, Z. Chen, F. Meng, and Y. Shi, "Pavement distress detection using random decision forests," in *International Conference on Data Science*. Springer, 2015, pp. 95–102.

[16] Y. Shi, L. Cui, Z. Qi, F. Meng, and Z. Chen, "Automatic road crack detection using random structured forests," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 12, pp. 3434–3445, 2016.

[17] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[18] J. Canny, "A computational approach to edge detection," in *Readings in Computer Vision*. Elsevier, 1987, pp. 184–203.

[19] Q. Zou, Y. Cao, Q. Li, Q. Mao, and S. Wang, "Cracktree: Automatic crack detection from pavement images," *Pattern Recognition Letters*, vol. 33, no. 3, pp. 227–238, 2012.

[20] H. Oliveira and P. L. Correia, "Crackit - an image processing toolbox for crack detection and characterization," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 798–802.

[21] T. S. Nguyen, S. Begot, F. Duculty, and M. Avila, "Free-form anisotropy: A new method for crack detection on pavement surface images," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, 2011, pp. 1069–1072.

ISBN 978-90-827970-1-5 © EURASIP 2018