Image Scale Estimation Using Surface Textures for Quantitative Visual Inspection

Ju An Park — Univeristy of Waterloo, ON, Canada
Chul Min Yeum — Univeristy of Waterloo, ON, Canada
Trevor Hyrnyk — Univeristy of Waterloo, ON, Canada
Email: {j246park, cmyeum, thrynyk}@uwaterloo.ca

Abstract

In this study, a learning-based scale estimation technique is proposed to enable quantitative evaluation of inspection regions. The underlying idea is that surface texture of structures (i.e. bridges or buildings) captured on images contains the scale information of the corresponding images, which is represented by pixel per physical dimension (e.g., mm, inch). This allows training a regression model that provides a relationship between surface textures on images and their corresponding scales. Deep convolutional neural network is used to extract scale-related features from the texture patches and estimate their scales. The trained model can be exploited to estimate scales for all images captured from structure surfaces that have similar textures. The capability of the proposed technique is fully demonstrated using data collected from surface textures of three different structures and achieves an overall average scale estimation error of less than 15%.

1 Introduction

With the recent advancements in vision sensors and computer vision algorithms, several data collection and feature extraction methods have been developed to enhance the vision-based inspection tasks of civil structures (i.e. bridges and buildings) in terms of accuracy and speed. These techniques typically involve collection and analysis of visual data involving structure component damage detection, localization, and quantification, of which comprehensive overviews are available in [1–3].

Despite tremendous effort made to advance visual inspection techniques through automated methods, its implementation in the industry remains limited. One of the factors inhibiting industry adoption is that damage detected on images cannot be quantitatively evaluate because their scales are unknown. We could utilize special equipment (i.e. stereo camera) or processes (i.e. including markers of known dimension in the inspection scene), but they incur either extra costs or time. Knowing the image scale (a pixel per length ratio) is important, as it allows for computation of physical dimensions of the detected structure damage and/or component.

To address this issue, a convolutional neural network (CNN) based image scale estimation technique is proposed to enable quantitative visual inspection by automatically estimating the scales of the collected images. The underlying idea of the proposed scale estimation is that surface textures convey scale information of the images. Thus, the surface texture to scale relationship can be learned through the use of a CNN. The estimated image scale (a pixel per physical dimension, in units, pixel/mm or pixel/in.) enables physical measurement directly from a single image. This technique would help enable quantitative automated visual inspection for existing visual inspection techniques.

2 Methodology

The proposed scale estimation technique aims to enable quantitative visual inspection using a single image. A key assumption for the proposed approach is that the scene of surface texture is unique at each image scale, and thus allows for their scale information to be determined. A CNN-based regression model is trained to extract the scales from surface textures. If users take images of scenes that include surface textures, which are similar to the one for training the model, the scales of the images can be estimated using the surface textures to measure the size of the inspection region on the collected images. This technique can be integrated into existing visual inspection algorithms that automatically detect areas of interest to enable a fully automated procedure that perform damage detection, localization, and quantification with the use of only images.

The overview of the technique is shown in Fig. 1. The technique is separated in two phases: model training and usability. For the first phase (model training), in step 1, surface texture scenes of

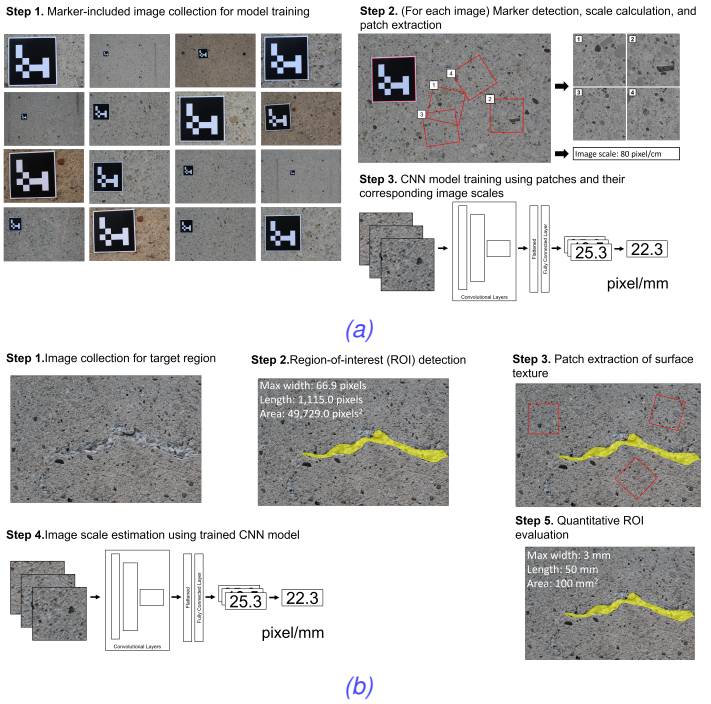


Fig. 1: Step-by-step procedures of (a) model training and (b) usability phases for image scale estimation.

the structure-of-interest are collected with the marker included at varying distances. In step 2, an existing marker detection algorithm [4, 5] is used to extract image scale and patches of surface texture rapidly and accurately from the image to form the training set. In step 3, the model is trained on the texture-image scale pairs.

Once the scale estimation model is trained, the technique can be utilized as shown in the usability phase. In step 1, the user takes images containing regions-of-interest (ROI) (i.e. spalling or cracking damages). In step 2, ROIs are manually or automatically detected. At this step, the ROIs can be quantified in terms of pixels. In steps 3 and 4, patches surrounding the ROI are extracted, fed into the model sequentially, and the estimated image scales averaged to compute the corresponding image. In step 5, the estimated image scale is used to quantify the detected ROIs by translating pixel area and length measurements found in step 2 into physical dimensions.

Note that since we assume that the entire region of the image has a single identical scale, the image should be captured parallel to the flat surface where damage is placed.

3 Experimental Setup

The technique is validated with respect to the following aspects:

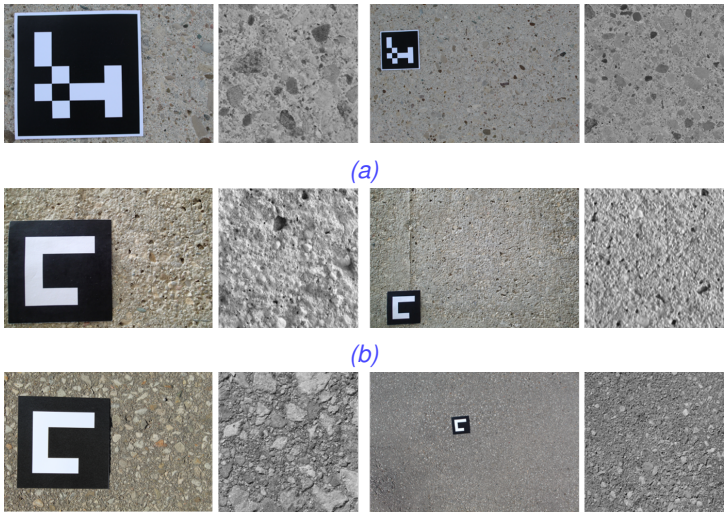
- relationship between image size of surface textures and the model accuracy; and
- algorithm robustness across various structures, each having different surface textures.

Note that all images unless otherwise mentioned are collected at a fixed focal length, and that the marker is also included in the testing images for validation purposes only, and is not required during the usability phase.

Images are collected from three different civil structures having different surface textures. These structures are located on the University of Waterloo's main campus in Waterloo, Ontario, Canada, as shown in Fig. 2. Many images of each structure are taken at various locations and distances. The distance is randomly chosen roughly from 0.5m to 2.5m. During data collection, the entire area of the marker is fully included in each image and the image is captured

(a) (b) (c)

Fig. 2: Overviews of structures with different textures (a) pedestrian bridge (PED), (b) building wall (BW), and (c) asphalt pavement (ASH)



(a) (b) (c)

Fig. 3: Two sample images and their corresponding patches, taken at a (left) close and (right) far distance for datasets (a) PED, (b) BW, and (c) ASH

in such way that its position on the image is randomly located. We collect the images relatively perpendicular to the scene. Examples of the three collected datasets are shown in Fig. 3.

A well-known CNN architecture, MobileNetV2 [6] is used as the base model with input size 299-by-299 while the top layer is configured for regression to enable image scale estimation from an input image, using mean average percentage error as the loss function, as shown in Eq. 1,

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} * 100 \quad (1)$$

where y_i and $\hat{y}_i$ is the actual and predicted image scale for a given patch, and n is the number of patches in a given set. Data augmentations such as random intensity changes (± 50), brightness changes ($\pm 20\%$), and horizontal and vertical flips were added. Stochastic gradient descent optimizer with learning rate, momentum, and stochastic decay of 10^{-4}, 0.9, and 0.01 were used.

The performance of the technique is examined in three different experiments, as follows:

- using patches with different sizes, 100X100, 350X350, and 850X850; and
- training and testing model performance using different unique surface datasets, PED, BW, and ASH.

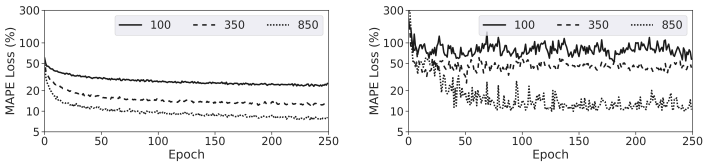
Prior to training the model, the dataset is split into training and testing sets where the model is trained on the training set and validated against the testing set. Note that to prevent leakage of training images into testing images and vice-versa, the datasets were split by scenes instead of by images. A single scene corresponds to a specific surface area on a structure, and contains several images taken at different distances of the same scene. Thus, by splitting the training and testing datasets by scenes, the model's performance can be assessed correctly. An overview of the collected datasets is shown in Table 1.

4 Experimental Result

The results of the experiments are shown in Fig. 4, 5 and Table 2. First, the effect of using different patch sizes are evident from Fig. 4, which shows MAPE loss curves for (a) training and (b) testing datasets. From the figure, it is obvious that larger patch sizes lead to the model converging at lower MAPE, since larger patches contain more texture information than do smaller patches which are more prone to error due to local variations in texture or lighting. Thus, it is recommended to use the largest possible patch size.

Table 1: Scale prediction results for all three structures using patch size 850X850: aggregated using either a mean or median function.

Datasets		*Total number of scenes (training/testing)*	*Total number of images (training/testing)*
	PED	22 (18/4)	191 (154/37)
	BW	14 (12/2)	434 (352/82)
	ASH	21(17/4)	182(149/33)

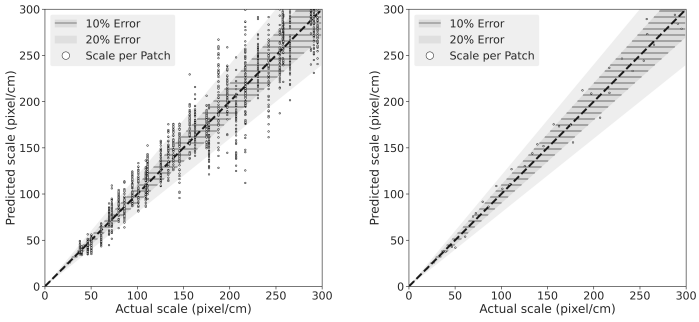


(a) (b)

Fig. 4: Effect of patch size on model accuracy, showing (a) Training and (b) testing loss curves on a log-linear plot for patch sizes: 100×100, 350×350, and 850×850 using the PED dataset.

An actual-to-predicted (AtP) scale scatter plot of PED testing dataset is shown in Fig. 5 which directly shows the performance of the model. In 5a, each point is the predicted image scale for a single patch, while 5b shows the median-aggregated image scale for a single image. The black dotted line follows the 1:1 ratio between actual and predicted scale which indicates a perfect model prediction. The inner-dashed band and the outer lightly coloured band are, in order, the 10% and 20% error margins. Only the median-aggregated plot is shown as no significant visual differences were observed between the two aggregation functions. It is easy to see that while the absolute value spread of the image scale predictions increase as the scale value increases, a majority of the model predictions stay within the 20% error margins.

The scale prediction results for all datasets are summarized in Table 2. The mean and median functions were used to aggregate several patch scales to a single image scale, and the values in the table are represented as mean values with their standard deviation and median values with their median absolute deviation. The performance of either aggregation is very similar to each other. Most of the model results perform on-average, with MAPE lower than 15%.



(a) (b)

Fig. 5: Actual-to-Predicted scale scatter (AtP) plots obtained from the PED training dataset using patch size 850X850: (a) scales for all patches, and (b) aggregated scales for each image using a median function. The black dashed line indicates a correct prediction and inner-dashed and outer-colored bands indicate 10 and 20 % error margins, respectively.

Table 2: Scale prediction results for all three structures using patch size 850X850: aggregated using either a mean or median function.

Datasets	Aggregation: *Mean*	Aggregation: *Median*
PED	$6.7\% \pm 4.0\%$	$7.3\% \pm 4.5\%$
BW	$15.8\% \pm 13.6\%$	$14.1\% \pm 11.9\%$
ASH	$10.5\% \pm 8.4\%$	$9.9\% \pm 8.3\%$

5 Conclusion

To enable quantitative assessment of detected ROIs, a CNN-based image scale estimation technique is developed. This technique estimates image scale from the unique texture information from the structure surface in images collected during inspection. A CNN model is trained to resolve visual surface textures to their corresponding image scales. Markers and relevant detection algorithms were used to automatically detect the marker in each image scene to estimate the image scale and extract patches from non-marker regions to form the ground-truth texture and image scale dataset. This dataset then is used to train the model, which, once trained, can be used to estimate scale for any image containing similar textures for quantitative evaluation of ROIs. The model performance is demonstrated using images collected from three different structures as training and testing datasets. On average, the model successfully estimates image scale solely by inferring from the surface texture, with less than 15 % error across all testing datasets.

Acknowledgments

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [RGPIN-2020-03979].

References

[1] M. R. Jahanshahi, J. S. Kelly, S. F. Masri, and G. S. Sukhatme, "A survey and evaluation of promising approaches for automatic image-based defect detection of bridge structures," vol. 5, no. 6, pp. 455–486, 2009. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/15732470801945930

[2] B. F. Spencer, V. Hoskere, and Y. Narazaki, "Advances in computer vision-based civil infrastructure inspection and monitoring," vol. 5, no. 2, pp. 199–222, 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2095809918308130

[3] X. Ye, T. Jin, and C. Yun, "A review on deep learning-based structural health monitoring of civil infrastructures," vol. 24, no. 5, pp. 567–585, 2019. [Online]. Available: https://doi.org/10.12989/SSS.2019.24.5.567

[4] S. Garrido-Jurado, R. Muñoz-Salinas, F. Madrid-Cuevas, and R. Medina-Carnicer, "Generation of fiducial marker dictionaries using mixed integer linear programming," vol. 51, pp. 481–491, 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0031320315003544

[5] F. J. Romero-Ramirez, R. Muñoz-Salinas, and R. Medina-Carnicer, "Speeded up detection of squared fiducial markers," vol. 76, pp. 38–47, 2018. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0262885618300799

[6] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2018, pp. 4510–4520. [Online]. Available: https://ieeexplore.ieee.org/document/8578572/