Automated Region-of-interest Localization and Classification for Visual Assessment

CHUL MIN YEUMS, JONGSEONG CHOI
and SHIRLEY J. DYKE

ABSTRACT

Visual assessment is a process to understand the state of a structure based on evaluations originating from visual data. Low-cost, high-performance vision sensors are providing new avenues for overcoming the spatial and temporal limitation in current human-based visual assessment when used in conjunction with aerial sensing platforms. However, past implementations are limited in their ability to deal with a high volume of images while only a small fraction of them are important for actual inspection. Such difficulty induces an unwanted high rate of false-positive and negative errors, reducing the trustworthiness and efficiency of their implementation. To overcome this challenge, we develop and validate a novel automated image localization and classification technique to extract regions-of-interest (ROIs) on each of images, which contain the target region of the structure for visual evaluation (TRI). First, ROIs are extracted based on the geometric relationship between the collected images and the TRIs using structure-from-motion algorithm. Second, unwanted ROIs corrupted by occlusion and image blur are effectively filtered by a robust image classification technique, called convolutional neural network. Then, a damage detection technique is applied only on such highly relevant and localized ROI images. The capability of the technique is demonstrated using a full-scale highway sign structure for the case of crack detection on weld connections.

INTRODUCTION

Changes in the appearance of a structure often provide obvious warning signs that a structure's condition is deteriorating. Thus, visual evaluation, the process of understanding the condition of a structure based on information that originates from visual data, remains the predominant means of assessing infrastructure systems as they gradually degrade over their lifetime.

Chul Min Yeum, Lyles School of Civil Engineering, Purdue University, West Lafayette, IN, 47907, USA.
Jongseong Choi and Shirley J. Dyke, School of Mechanical Engineering, Purdue University, West Lafayette, IN, 47907, USA.

Currently, visual evaluation procedures require that human engineers are on-site and actively involved in at least one or more of the steps, including observation, data collection, data analysis, or decision-making. However, such human-oriented visual evaluation does have certain limitations. First, it is expensive and time-consuming. Civil structures are often large and complex, and can be placed in a harsh environment. Thus, there are often challenges in reaching and accessing many critical regions for viewing the structure. Expensive equipment such as a scaffolding is often required to reach hardly accessible structures, although it can cover the only small region with a limited degree-of-freedom. Moreover, such equipment requires training and is usually operated by specially trained staff members.

In this study, region-of-interest (ROI) localization and classification techniques are developed and incorporated to enable realization of automated vision-based visual assessment. An ROI is the portion of an image that contains the region of the structure that is targeted for visual evaluation, which is denoted as the targeted region for inspection (TRI). Damage sensitive components or areas of a structure would be assigned as the TRIs. The ROI on each image is first automatically localized based on the geometric relationship between the collected images and TRIs using structure-from-motion algorithm (SfM) [1]. However, depending on the viewing angle, the ROIs in certain images may not be favorable for visual evaluation due to occlusions that impede the view of the TRI. A robust image classification technique, using a convolutional neural network algorithm, is further implemented to eliminate the useless ROIs. With such a binary occlusion classifier (BOC), only the valuable portions of certain images are retained for efficient and reliable vision-based visual inspection. The developed technique is demonstrated through performing visual inspection of the welded connections in a full-scale highway truss structure.

The major contribution of this study is to develop a technique to facilitate the real application of existing damage detection algorithms with large volumes of actual images collected using a commonly available camera and sensing platforms, such as a UAV. Large numbers of images or video may readily be acquired for visual inspection purposes. However, despite the enormous potential of this technology, automated processing is just not possible at this time. Manual viewing, sorting, and analysis is costly and inconsistent and prone to human error. Past research has been successful in automating the analysis of individual images for damage detection. However, to apply these well-established methods, preselected and localized images are needed that are known to contain a region of interest. Instead, the developed technique begins by searching a large set of images to extract damage sensitive areas. By detecting and extracting these regions (ROIs) from many different viewpoints, the detectability of damage can be dramatically increased even if it is small, and false-positives can be reduced by limiting and regularizing the search areas. This technique will be a key enabler in the automated visual evaluation, breaking down the existing barriers that have impeded the use of large volumes of complex images.

OVERVIEW OF THE TECHNIQUE

The technique is intended to achieve the visual assessment scenario proposed by the authors [2]: A UAV equipped with a high-resolution camera arrives at a candidate structure. Following a flying path designed *a priori*, the UAV automatically flies using GPS, collecting and recording images near designated target areas from many viewpoints. Using those images, processing takes place on the large volume of images, and damage present in the structure is detected, localized, and quantified automatically without requiring the involvement of human inspectors. This information would provide evidence to facilitate better decision-making related to repair and maintenance priorities for other structures.

A key step that enables this scenario is to extract highly relevant and useful ROIs corresponding to the TRIs on the collection of complex images, supporting robust and efficient visual evaluation. This step is needed because, regardless of the flying path designed *a priori*, each image will still include a large portion of irrelevant areas. Uncertainty in GPS data, occlusions due to the complex geometry of the structure, and generally, the capabilities of photography. Thus, both irrelevant images and irrelevant portions of images must be filtered out in advance before implementing specific damage detection methods to avoid unwanted erroneous results.

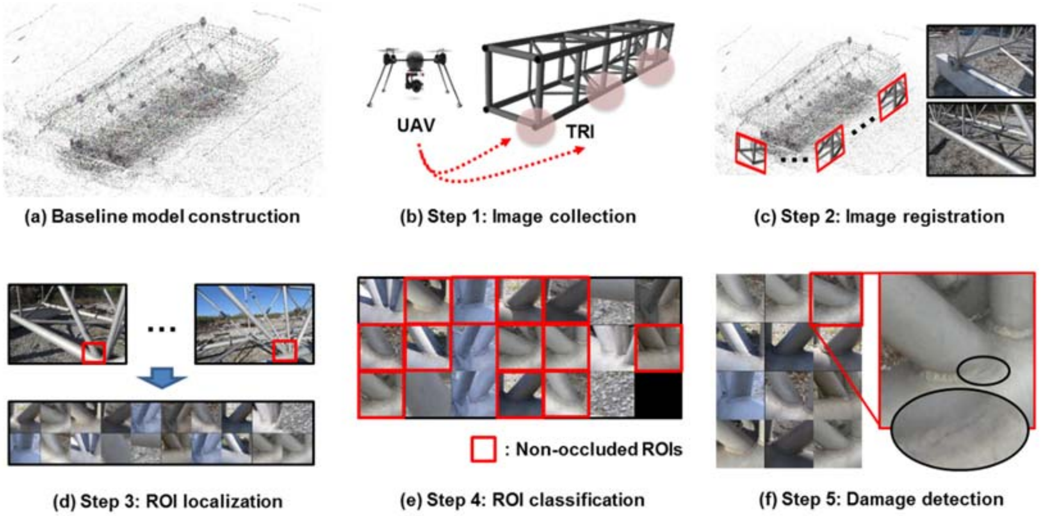

UAV
TRI
(a) Baseline model construction
(b) Step 1: Image collection
(c) Step 2: Image registration
: Non-occluded ROIs
(d) Step 3: ROI localization
(e) Step 4: ROI classification
(f) Step 5: Damage detection

Figure 1. Overview of the technique developed: (a) construct a baseline model of target structure (pre-processing), (b) collect images around each of the targeted regions for inspection (TRIs), (c) register the images into the baseline model, (d) localize regions of interest (ROIs) on the images, (e) classify the favorable ROIs, and (f) detect (crack) damage on the ROIs.

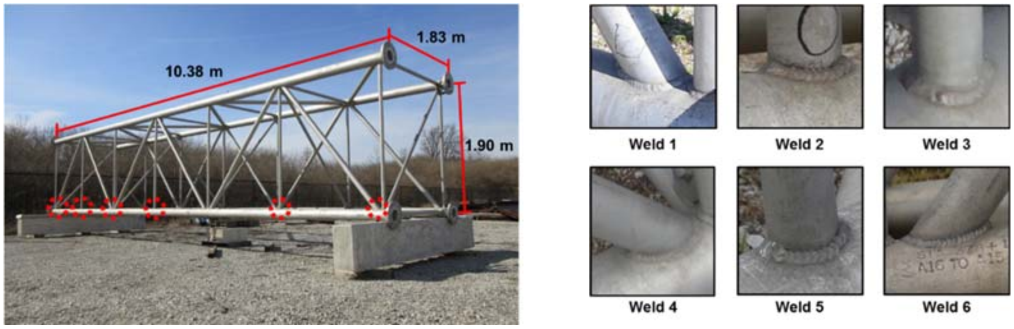
The proposed approach first requires building a baseline 3D model of the target structure, in advance, using a large number of images (hereafter, baseline images) that cover the entire structure. This process is carried out only one-time at the beginning. Then the resulting model will subsequently be utilized for visual inspection with newly collected images (hereafter, test images) during each inspection task over the years. During an inspection, a set of test images is collected around each TRI, and the ROIs on test images are localized by geometrically mapping the ROIs to the corresponding TRI. An image classification technique is further implemented to identify and extract the ROIs that are favorable for damage detection. Finally, a method to identify the damage type of interest is evaluated on only these highly relevant and visible ROIs. Each damage identification method would then be applied to each ROI found to be relevant.

In Fig. 1(a), as a preliminary step, a baseline model is constructed using SfM from a large volume of images collected around a target structure. The baseline model consists of 3D points with descriptors and calibrated images [3-4]. The 3D points in this model and their descriptors are exploited to register the test images into this model. The TRI is defined by assigning new 3D point(s) in the baseline model through 2D point match on the calibrated images [3].

For actual visual evaluation, the test images are first collected from near each TRI in the structure, as shown in Fig. 1(b) and mentioned in the proposed visual evaluation scenario. The red blurred circles in the diagram indicate the selected TRIs for this structure. A large volume of images are first collected from different viewpoints and locations around each TRI. Next, in Fig. 1(c), the collected images are registered into the baseline model by matching the 3D points in the baseline model with the 2D features on the test images using their descriptors. Then, external (location and orientation) and internal camera parameters in each of test images are calibrated in the coordinate system of the baseline model. The geometric relationship between test images and each TRI in the baseline model is identified. Third, in Fig. 1(d), each ROI is extracted by mapping of the TRI on to each of the test images. Since the size of the ROI is computed by geometrically projecting the TRI on the image, the scale (size of the TRI in the ROI) is identical, improving the performance of damage detection algorithms. However, the visibility of some portions of the ROIs may be hindered by occlusions due to the complex geometry of the structure. To filter out such undesirable occlusion ROIs, a robust image classification algorithm, based on convolutional neural network algorithms, is implemented in Fig. 1(e). This binary occlusion classifier (BOC), is used to distinguish between the non-occluded and occluded ROIs, and is trained using manually labeled ROIs obtained from the baseline model. Then, the trained BOC is applied to the ROIs extracted from test images. Finally, damage on the TRI is detected on the classified ROIs in Fig. 1(f).

EXPERIMENTAL VERIFICATION

Description of Experiment



(a) (b)

Figure 2. Description of a full-scale highway sign truss structure: (a) dimensions of the structure, and (b) six welded connections defined as the TRIs, which are marked as dotted red circles in (a).

For demonstrating the method developed here, a full-scale highway sign truss structure (hereafter, the test structure) is used, shown in Fig. 2(a). This structure was originally built for supporting highway message signs. Currently, it is located outside the Bowen Large-scale Structural Engineering Laboratory at Purdue University. The test structure is composed of six cubic segments. It has four main chords, twenty-eight vertical braces, twenty-four diagonal braces, and seven internal braces oriented diagonally to the main chord. All members are tubular sections. The diameters of the main chord, vertical brace, and diagonal (including internal) brace are 152.4, 63.5, and 76.2 mm, respectively. The braces are connected to the main chords using welds.

Six fillet welded connections between the vertical or diagonal braces and main chords, which are regarded as critical components, are selected as TRIs for the test structure. As shown in Fig. 2(b), since tubular braces and tubular chords are connected, the shape of the weld is a warped circular line. The dotted red circles represent the TRIs. From right to left in Fig. 2(a), these are denoted as Weld 1 to Weld 6. Welds 2, 3, and 5 (denoted as Type 1) and Welds 1, 4, and 6 (denoted as Type 2) are created where the vertical and diagonal braces are attached to the main chord with diameter 63.5 and 101 mm, respectively.

Baseline Model Construction

To construct the baseline model for this verification, a total of 5,321 baseline images were collected from the test structure over five months. Images were collected on 11 different days at different times of day and/or weather conditions. Sample baseline images are shown in Fig. 3. Each column in this figure shows images captured on different days. Having such a large collection of images under various environmental conditions enables matching robust and unique features, that are invariant to environment changes, and have enhanced descriptors that normalize the variance resulting from varying lighting conditions. We do not fly a UAV, although image acquisition is designed to be representative of the process needed for practical use of this technique using UAVs. Thus, it is important to note that these images are acquired from locations that are not pre-determined.

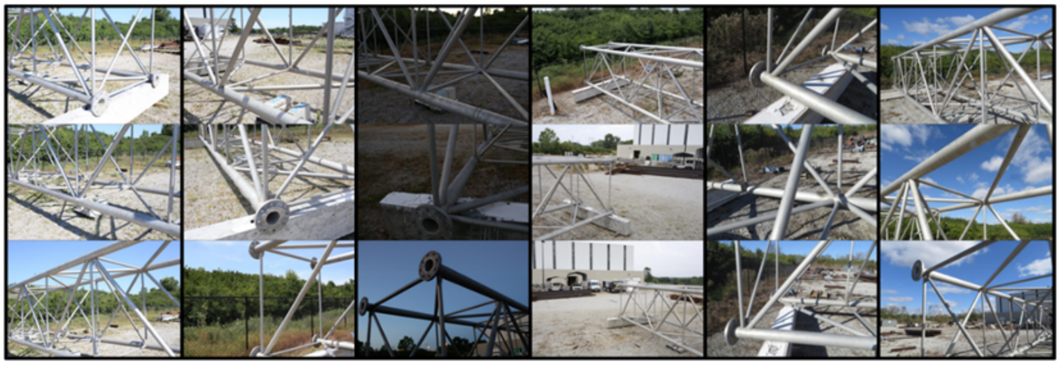


Figure 3. Sample baseline images used for constructing a baseline model: Images in each column are collected on different days and under different lighting sand weather conditions.

We use VisualSfM to compute the projection matrices and to construct the 3D point cloud. VisualSfM is a non-commercial, free software having a user-friendly graphic user interface (GUI) [5]. To define the TRIs in the baseline model, we manually attach red circular stickers at each weld, and these are clearly visible on the images. Four stickers are attached at the ends of both major and minor axes if the cross-section of the welds is assumed to be an ellipse (precisely, they are warped ellipses). Among the sets of test images collected from 11 different days, only three image sets were captured from the test structure after attaching these stickers. By matching the same sticker in a few different images (precisely, the center location of the sticker on the image), the 3D location of the corresponding sticker can be computed. Once the 3D locations of the two ends of the major axis for the TRIs are computed, the 3D center location and diameter (length of the major axis) of the weld can be computed. These two parameters can then be used for defining a virtual sphere [1]. Then, we apply an augmentation factor of 1.5 to the diameter of the virtual sphere to ensure we include the entire TRI when extracting each ROI.

Next, the ROI are extracted from the baseline images to train the BOC for filtering out occluded ROIs. The ROIs extracted from the baseline images are only used for training the classifier, and this step has nothing to do with the actual visual inspection. Also, to be useful for visual inspection, an ROI must be sufficiently visible in the image. In this study, the ROIs with a pixel size smaller than the diameter of the virtual sphere in the corresponding to TRI (in mm) are automatically removed. For example, if the minimum dimension (width or height) of an ROI corresponding to a Type 1 connection is smaller than 95.25 pixels (63.5 x 1.5 (augmentation factor)), that ROI is rejected. Thus, both the ROIs localized from the baseline images, and the test images for Type 1 connections are always larger than 95.25 pixels.

Knowing the projection matrix in each baseline image, and the 3D location of the TRIs in the baseline model, the ROIs of all six TRIs are localized from the baseline images. Since the images are collected from various angles and positions, the ROIs extracted show various viewpoints of the TRIs, but their scale is almost identical. For example, when the image is captured close to the TRI, the size of the ROI is large.

A total of 4,298 ROIs corresponding to all six welds are localized from the baseline images. We manually annotate these images to construct a dataset for training the BOC. Non-occluded ROIs, denoted as positive, are defined as those in which the entire weld line on the ROI, that can be maximally viewed at the corresponding image location, is not interrupted by any object(s) in front. The remainder of the ROIs are annotated as negative for training. Among 4,298 ROIs, 945 ROIs are annotated as negative, and the rest are positive. All labeled ROIs are randomly divided into 2,144 (50%), 1,077 (25%), and 1,077 (25%) images for training, validating and testing. Note that the samples are used here to evaluate the performance of the trained BOC, and not actual use for extracting the ROIs on the test images. We implement a popular ImageNet CNN model called Alexnet, framed in MatConvNet library [6].

The BOC successfully attains a relatively high accuracy. We obtain rates of 89.73% (743/828 images) true-positive (true classification of non-occluded ROIs) and 91.83% (225/245 images) true-negative, respectively. The precision is 97.37%, defined as the number of true-positives over the total number of positives. Although these rates will vary slightly depending on the CNN architecture used and its parameters, the overall performance of this approach is quite successful. These classification results imply that the trained BOC can successfully filter the occluded ROIs with high accuracy.

ROI Localization and Classification

The number of images collected from all six of the TRIs is listed in Table I. First, SIFT features and descriptors are extracted from each test image at the original resolution. Second, these features are matched with 3D points in the baseline model to register each of the test images. Instead of exhaustive searching for the closeness of the descriptors, we used a GPU-implemented k-nearest search algorithm in MATLAB to rapidly find the best pairings. The criterion to accept matches is the same as the one implemented in vl_ubcmatch in the VLFeat library: “A descriptor D1 is matched to a descriptor D2 only if the distance between D1 and D2 multiplied by 1.5 is not greater than the distance of D1 to all other descriptors" [7].

TABLE I. RESULTS OF THE ROI LOCALIZATION AND CLASSIFICATION

	Weld 1	Weld 2	Weld 3	Weld 4	Weld 5	Weld 6
# of images	119	77	88	84	60	55
# of localized ROIs	104	51	54	70	45	47
# of classified ROIs (positive/negative)	69/35	49/2	48/6	47/23	44/1	33/14
Precision	92.75%	100%	97.92%	85.11%	100%	90.91%

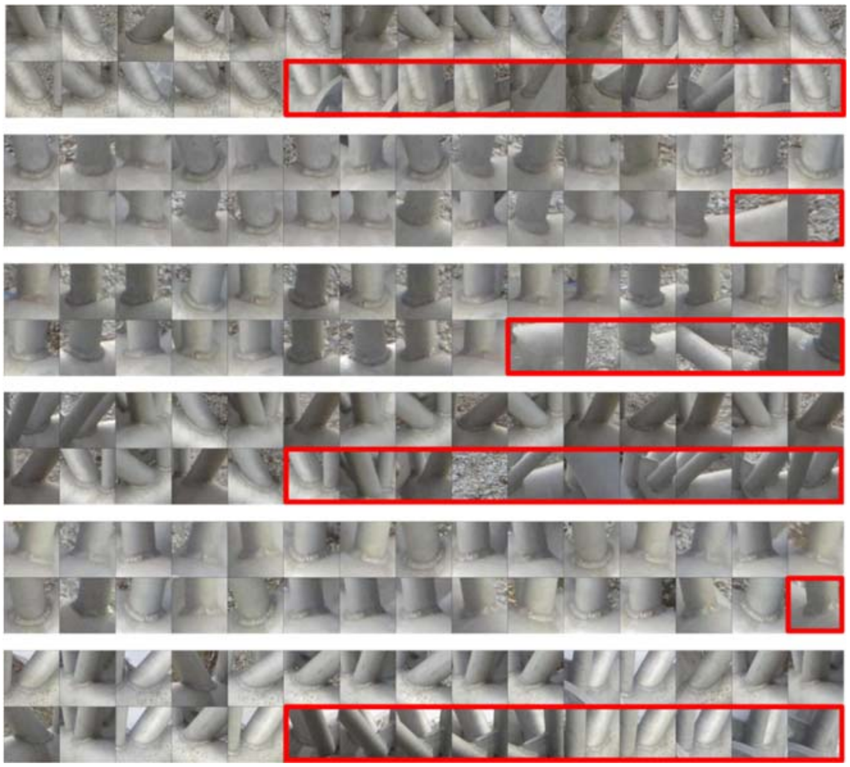


Figure 4. Examples of ROIs that have been localized and classified from the set of test images: Each set of 30 localized ROIs corresponds to Welds 1 to 6 (from top to bottom). A maximum of 10 negatively classified ROIs, are positioned at the end of the set (marked with a red box). The rest are classified as positive.

By the pairing between the 3D points in the baseline model and the 2D features in each test images, the projection matrices and one lens distortion parameter are calibrated. Based on the projection matrix associated with each test image, the ROI localization and classification outcomes are shown in Fig. 4. In each double row, 30 random localized ROIs are shown for Welds 1 to 6 from top to bottom. The actual number of localized ROIs is shown in Table I. Because any ROIs having insufficient resolution and not including the ROIs are rejected, the number of ROIs identified in these results are fewer than the total number of test images. Next, the trained BOC is applied to the localized ROIs to filter out the occluded ROIs that are not useful for visual evaluation. Among the 30 ROIs shown in Fig. 4, the ROIs that are classified as negative (with a maximum of 10 shown), are added at the end of the list and marked with a red box. The rest of them are classified as positive. The total number of ROIs that are classified as positive and negative is listed in Table I. Also, after manually annotating the classified ROIs extracted from the test images, the precision in classification is computed. Overall the performance of this classifier is successful based on its high precision. Such high precision, which is the number of true-positives over the number of positives, implies that when reliable damage detection methods are applied to these ROIs, they are unlikely to produce false-positive errors because the most of the positive ROIs are true-positive (non-occluded) ones.

CONCLUSION

This study presents and experimentally verifies an automated technique for ROI localization and classification that will directly enable robust, vision-based, visual structural assessment. This technique overcomes a major practical barrier in the use of large volumes of complex images, such as those collected with UAVs for assessing structural condition. A key technical achievement in this study is to make the best use of the collected images to (1) efficiently localize ROIs by computing the 3D geometric relationship between the TRIs and the images using SfM, and (2) obtain the most useful ROIs by learning their 2D unique visual patterns using convolutional neural network algorithms to implement a BOC. Ultimately, we expect that this technique will be a major enabling method needed to automatically assess various large-scale civil structures using the images collected from UAVs.

ACKNOWLEDGEMENT

Partial support from the National Science Foundation is acknowledged under NSF Grant No. CMMI-1645047. We acknowledge the NVIDIA Corporation for the donation of a high-end GPU board.

REFERENCES

1. Yeum, C. M., Choi, J., & Dyke, S. J. (2017). Autonomous image localization for visual inspection of civil infrastructure. Smart Materials and Structures, 26(3), 035051.
2. Yeum, Chul Min, & Dyke, S. J. (2015). Vision-Based Automated Crack Detection for Bridge Inspection. Computer-Aided Civil and Infrastructure Engineering, 30(10), 759–770.
3. Hartley, R. I., & Zisserman, A. (2004). Multiple View Geometry in Computer Vision (Second). Cambridge University Press, ISBN: 0521540518.
4. Snavely, N., Seitz, S. M., & Szeliski, R. (2008). Modeling the World from Internet Photo Collections. International Journal of Computer Vision, 80(2), 189–210.
5. Wu, C. (2013). Towards linear-time incremental structure from motion. In 3D Vision-3DV 2013, 2013 International Conference on (pp. 127–134). IEEE.
6. Vedaldi, A., & Lenc, K. (2014). MatConvNet - Convolutional Neural Networks for MATLAB. arXiv:1412.4564.
7. Vedaldi, A., & Fulkerson, B. (2010). VLFeat: An open and portable library of computer vision algorithms. In Proceedings of the 18th ACM international conference on Multimedia. 1469-1472.