9th International Conference on Structural Health Monitoring of Intelligent Infrastructure
2019 ST. LOUIS SHMII-9
AUGUST 4-7, 2019 | St. Louis, Missouri USA
Organized by MISSOURI S&T

Image Data Analytics to Support Engineers' Decision-Making

Tianyi Zhao[1], Zhaozheng Yin[1], Ruwen Qin[2], and Genda Chen[3]
[1]*Dept of Computer Science,* [2]*Dept of Engineering Management and Systems Engineering*
[3]*Dept of Civil, Architectural and Environmental Engineering*
Missouri University of Science and Technology, USA
Emails: {tznbz, yinz, qinr, gchen}@mst.edu

Abstract

Robots such as drones have been leveraged to perform structure health inspection such as bridge inspection. Big data of inspection videos can be collected by cameras mounted on drones. In this project, we develop image analysis algorithms to support bridge engineers to analyze the big video data. Bridge engineers define the region of interest initially, then the algorithm retrieves all related regions in the video, which facilitates the engineers to inspect the bridge rather than exhaustively check every frame of the video. To perform this task, we propose a Multi-scale Siamese Neural Network. The network is initially trained by one-shot learning and is fine-tuned iteratively with human in the loop. Our neural network is evaluated on three bridge inspection videos with promising performances.

1. Introduction

Traditionally, performing bridge inspections in hard-to-access areas is disruptive, difficult and dangerous. In many cases, bridges must be closed to traffic and inspectors must be lifted by heavy equipment. Manual inspection is time-consuming and costly. Using robotics to conduct bridge inspections will be safer, faster, and cheaper. Currently, big data from bridge inspections can be collected from videos recorded with cameras mounted on drones. With a frame rate of 30 frames per second, 108,000 frames can be recorded in one hour. It is a tedious and inefficient process for bridge engineers to watch hours of video footage for bridge inspection.

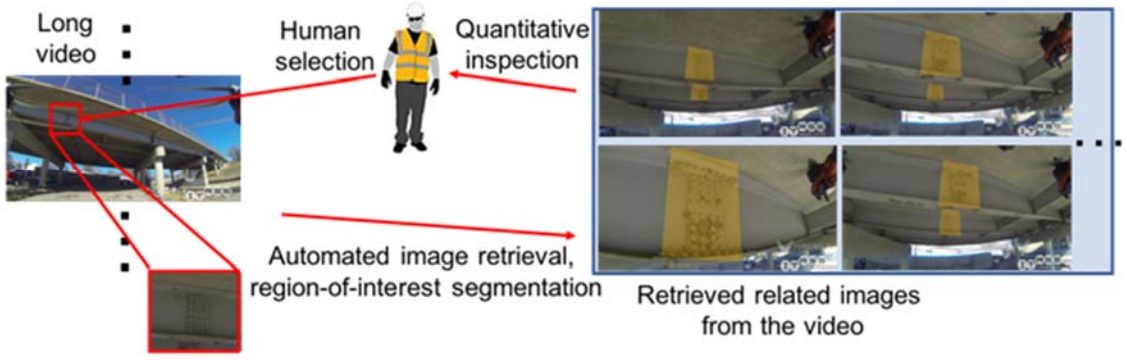

Fig. 1 Automatic retrieval of the region of interest based on the initial input by inspectors.

This project aims to deploy image analysis methodologies to provide decision-making support for bridge inspection through long videos. Fig. 1 illustrates the main steps of an automatic retrieval of the region of interest from a long video. An inspector first selects some regions of

interest (e.g. joints, beams, and surface) in a frame. The image retrieval algorithm developed in this project then finds all related frames in the video. Finally, the collected set of images with localized regions of interest can be evaluated automatically by computer algorithms or verified by inspectors.

The main challenges include: (1) the viewpoint is changing within a video captured by a camera mounted on a drone, (2) the camera vibration introduced by the drone movement affects the image quality, (3) the regions of interest have different scales in the videos, and (4) the regions to be inspected by bridge engineers may have different visual appearance or types.

A simple template matching or comparing the similarity between hand-crafted features of the query image and reference images may not overcome the previous challenges. Neural networks were used from 1950s to solve the supervised learning problem. At the end of the 20th century, neural networks were applied to the handwriting digital recognition task and achieved superior performance. The neural network method relies on big training data, efficient optimization methods and powerful computation resources. In 2012, deep neural networks [1] were proposed to solve the large scale image classification problem. Since then, deep neural networks remained the hottest machine learning topic in many industry applications.

In this project, after a bridge engineer selects a target object, we aim at retrieving the similar object from every frame of the video. The goal of our project is to assist engineers with less human effort (e.g., selecting the region of interest by a single image cropping operation). The proposed system has three main contributions: (1) we propose a Siamese neural network that extracts features from the target object patch and video frame using the same network architecture and detects the region of interest by feature similarity comparison; (2) we extend the network into a Multi-scale Siamese Neural Network, which is able to detect the region of interest at multiple scales when the camera on a drone moves away or towards the bridge; (3) since we only have one training sample from the initial selection of the bridge engineer, we leverage the one-shot learning to fine-tune the pre-trained network to the bridge inspection domain, and we propose an iterative approach to further refine the network performance with human-in-the-loop.

2. Methodologies

2.1 Preliminaries on Convolutional Neural Network

The convolutional neural network (CNN) is composed of a large amount of neurons organized by multiple convolutional layers. Each layer contains multiple neurons. Each neuron has one convolutional kernel that can perform one particular task (e.g., detecting one particular pattern). From the perspective of transformation function, the first layer transfers the input image to a stack of feature maps. Then the following layers continue to transfer the feature maps to more abstract feature maps. The lower level feature map is more local, for example, edge or texture pattern recognition. The higher level feature map is more abstract, for example, part or object detection. In addition to convolutional layers, CNN has some other layers including pooling layer, normalization layer, fully connected layer, and different connections between layers (e.g., skip connection, dense net, split and merge, multi-scale Inception).

2.2 Siamese Neural Network

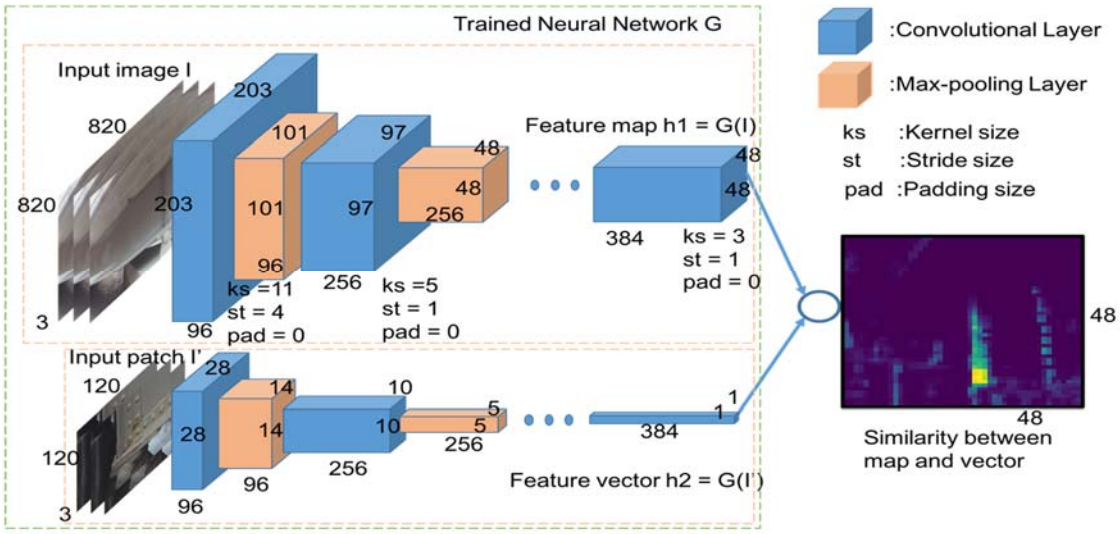

Fig. 2 The proposed Siamese neural network.

The Siamese neural network [2] contains two network architectures which share the same network architecture to compare two images with the same size. We propose a new Siamese neural network that can compare two images with different sizes (i.e., the target object patch and the test image). Our network architecture contains mainly convolutional layers since fully convolutional layers [3] can adapt to input images with different sizes and generate the output with the corresponding size. As shown in Fig. 2, the channel number (or the number of convolutional kernels) increases at each layer, while the size (width and height) of the feature maps decreases at each layer. The max pooling layer, whose stride size is 2, decreases the feature map size by 2, as illustrated by one toy example of a single slice of the feature map in Fig.3. The size of the feature map generated from a convolutional layer follows the equation:

$$W_O = (W_I + 2 \times pad - ks) / st + 1, \quad (1)$$

where W_O and W_I denote the size of output and input, respectively. *pad*, *ks* and *st* denote the number of rows for zero-padding, kernel size and stride size, respectively. For example, the first convolutional layer of the test image in Fig. 2 has $pad = 0, ks = 11$, and $st = 4$, so $W_O = \frac{820+2\times0-11}{4} + 1 = 203$.

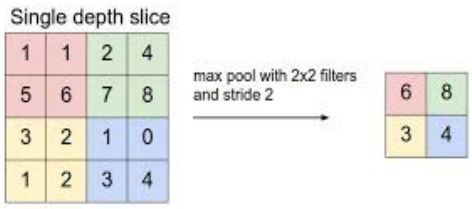

Fig. 3 The max-pooling example.

The features from the test image *I*, $h1 \in R^{W\times H\times K}$, are extracted by Siamese neural network *G*, where *W*, *H* and *K* denote the width, height and channels of the feature map, respectively (*W*=48, *H*=48, *K*=384 in Fig. 2). The feature maps of the test image contain the feature vectors at every location, i.e. $h1_{w,h}$ is the feature vector at location (*w*,*h*). Since the fully convolutional layer can accept different input sizes, given the target object patch, one single feature vector, $h2 \in R^{1\times1\times K}$, is extracted by the same neural network *G*, The similarity between two feature vectors is measured by

$$P(h1_{w,h}, h2) = Sigmoid\left(\frac{h1_{w,h}\, h2}{|h1_{w,h}||h2|}\right). \quad (1)$$

where $Sigmoid(x) \triangleq 1/(1 + x^{-1})$. The similarity computation between the target object patch and every location in the test image provides a 2D probability map that tells us how likely the

object is detected at specific locations in the test image. The two shared network architectures (G) can be trained in an end-to-end manner.

2.3 Multi-scale Siamese Neural Network

During the inspection, the camera on a drone may move towards or away from the target area, which changes the object scale continuously. Our image-patch Siamese Neural Network in Fig.2 can work well at one scale, but it may fail if the scale changes too much. Thus, we propose a multi-scale Siamese neural network as shown in Fig.4. We up-sample and down-sample the target object patch to a few scales (e.g. $W1 \times H1$ and $W2 \times H2$ in Fig.4). The smaller patch I''

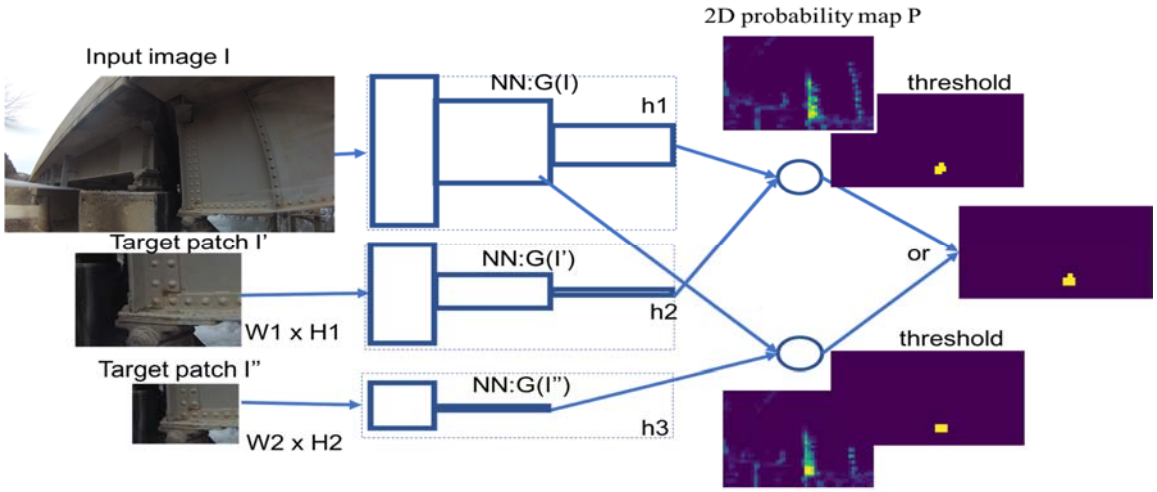

Fig.4 Multi-scale Siamese Neural Network.

with size $W2 \times H2$ (the camera moves far away from the bridge) will generate the feature vector, $h3 \in R^{1\times1\times K2}$, at the lower level of the network G. The larger patch I' with size $W1 \times H1$ (the camera moves towards the bridge) will get the feature vector, $h2$, at the higher level of the network G. The test image will also be given to the network G and generate the feature maps at different levels. At each level, a 2D probability map is generated, as described in the previous section. The generated probability map is turned into a bitmap through a threshold operation. The overall prediction is the union of the thresholded results from all scales.

2.4 One-shot Learning and Fine-tune with Human in the Loop

The region of interest initially selected by bridge engineers, as a single image patch, is obviously not enough for training a neural network from scratches. Thus, we deploy the pre-trained neural network model Alexnet [1] and then fine-tune the network to fit our bridge inspection project. The Alexnet is trained from the large-scale ImageNet [4] dataset. This dataset contains 1000 categories, and each category contains more than 1000 images. Those 1000 categories don't include the region of interest in our bridge inspection project, but the images in the dataset contain similar texture patterns. To let the network fit the bridge inspection problem much better when we only have one (or a few) training samples, one-shot (or few-shot) learning [5] is intuitively suitable for this situation.

We treat the region of interest detection as a binary classification problem. The region of interest cropped by human is the positive sample, then on the same frame, the other pixels belong to the background. Accordingly, the ground truth map Y is generated based on the positive and negative samples. Let P denote the detected probability map, then the loss to be minimized is a weighted cross-entropy function:

$$Loss = \sum_{w,h} -Y_{w,h}\, log\big(P_{w,h}\big) - \alpha\big(1 - Y_{w,h}\big) log(1 - P_{w,h}). \quad (2)$$


Fig.5 Mask overlays on image.

Since there are more negative pixel samples than positive samples, we use weight α to balance the two classes (α is set as 10 in our experiments). The loss function is calculated at each level of our Multi-scale Siamese Neural Network. The overall loss function is a weighted sum of the loss from all levels. After the one-shot training from the initially labeled sample, the detection results over the whole video are generated, which can be visualized as a mask overlaid on the original image as shown in Fig.5. Bridge experts can quickly skim the results and identify some false positives (incorrect detections) and select some correct detections with large appearance variations. Through the interaction with small human efforts, we can have a little more training data to fine-tune the Multi-scale Siamese Neural Network. The process can be iterated until satisfied detection performance is achieved. In our experiments, we manually select 15 frames in each iteration, which are added into the fine-tune process. All the frames selected from the current and previous iterations are used for fine-tuning the network in the current iteration.

3. Experiments

The feature extraction network in our Multi-scale Siamese Neural Network is similar as the first 5 layers of Alexnet. The kernel sizes for the 5 layers are 11, 5, 3, 3, 3. The numbers of kernels for the 5 layers are 96, 256, 384, 384, and 256. The stride size is 4, 1, 1, 1, and 1. The patch sizes for multi-scale Siamese neural network are 120,70, and 50, and the feature vectors are extracted from the 5th, 3rd, 2nd convolutional layer. The optimizer is Stochastic Gradient Descent. The learning rate is set as 10^{-1} , then decreases by 10 until 10^{-4} when the loss doesn't decrease.

Videos	*One-shot learning*			*Iteration 1*			*Iteration 2*		
	prec	recall	f1	prec	recall	f1	prec	recall	f1
Video 1 (7m:25s)	.333	.448	.382	.730	.589	.652	.877	.514	.673
Video 2 (5m:42s)	.376	.75	.501	.693	.964	.806	.678	1.0	.809
Video 3 (2m:26s)	.348	.712	.467	.773	.787	.78	.818	.863	.84

Table. 1 The detection performance on 3 videos. One shot learning column shows the results generated by the model trained by the first image patch only. Iteration 1 shows the results after the first round of human interaction with 15 frames. Iteration 2 shows the results after the second round of human interaction with 30 frames which include the 15 frames from the first round. 'prec' denotes the precision. 'f1' denotes the f1 score.

Our network is evaluated on three bridge inspection videos. We manually label the ground truth for each frame of the videos. The number of target object (bridge joint in the experiments) in three videos is 181, 56, 70, respectively. For each video, we perform two iterations of human

interaction. Table 1 summarizes the evaluation results, from which we observe that iteratively fine-tuning can improve the performance gradually. Due to the large appearance variation, there are still some miss detections and false alarms by our network using the initial one-shot training. More training samples will definitely help overcome this problem and we leave this as our future work. Some qualitative results are shown in Fig. 6. The first column is the hard prediction, which is thresholded form the soft prediction (the second column of Fig. 6). The last column is the original test image with the bounding box representing the detected area.

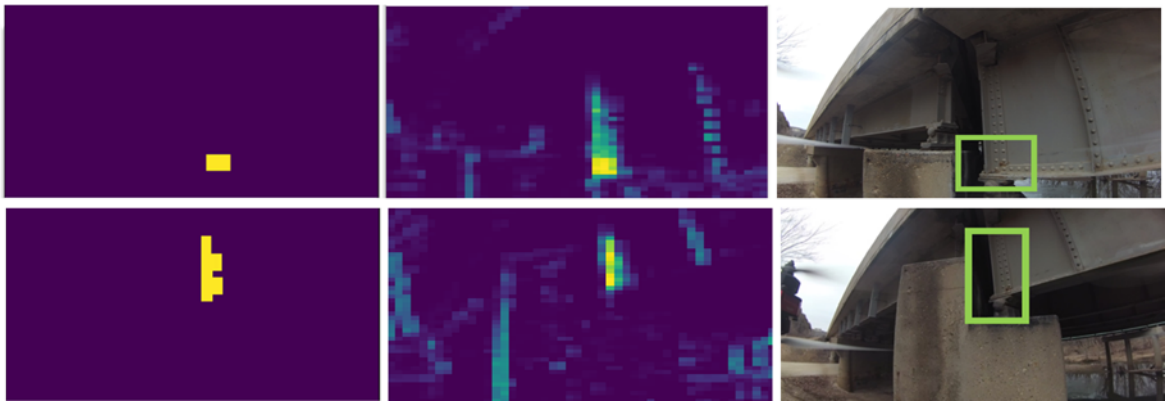

Fig. 6 Hard prediction, soft prediction and detected bounding box on the test image.

4. Conclusions

In this paper, we aim to develop image analysis algorithms to provide decision-making support for bridge inspection through long videos. Our proposed algorithms include: (1) image-patch Siamese neural network; (2) multi-scale Siamese neural network; (3) one-shot learning and iteratively fine-tuning the pre-trained network with human-in-the-loop.

5. Acknowledge

We would like to thank the financial support provided by INSPIRE UTC, and CS and EMSE department of MST, and the video data of bridge Inspection provided by Dr. Genda Chen.

References

[1] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." *Advances in Neural Information Processing Systems*. 2012.
[2] Koch, Gregory, Richard Zemel, and Ruslan Salakhutdinov. "Siamese neural networks for one-shot image recognition." *ICML Deep Learning Workshop*. Vol. 2. 2015.
[3] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2015.
[4] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2009.
[5] Cai, Qi, et al. "Memory Matching Networks for One-Shot Image Recognition." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.