Automation in Construction 106 (2019) 102849

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon

Underground sewer pipe condition assessment based on convolutional neural networks

Syed Ibrahim Hassan[a], L. Minh Dang[a], Irfan Mehmood[a], Suhyeon Im[a], Changho Choi[b], Jaemo Kang[b], Young-Soo Park[b], Hyeonjoon Moon[a,*]

[a] Department of Computer Science and Engineering, Sejong University, Seoul, Republic of Korea
[b] Korea Institute of Civil Engineering and Building Technology (KICT), Ilsan, Republic of Korea

ARTICLE INFO

Keywords:
Deep learning
Closed circuit television (CCTV)
Convolutional neural network
Automation
Sewer assessment
Text recognition
Maximally Stable Extremal Regions (MSER)

ABSTRACT

Surveys for assessing the condition of sewer pipeline systems are mainly based on video surveillance or CCTV, which is a time-consuming process that relies heavily on human labor because an operator has to watch videos, looks for defects and decides the defect's type manually. Previous research required suitable handcrafted features that were inefficient in analyzing sewer pipeline condition, so a robust and efficient framework is crucial as it eliminates the time-consuming tasks and helps the operator access condition of sewer systems more efficiently. This study proposes a defect classification system on CCTV inspection videos based on convolutional neural networks (CNN). The dataset was manually constructed and validated by extracting the images from CCTV videos, and the images were labeled according to six predefined defects. The CNN model was fine-tuned before training, and trained on a total of 47,072 images (256 × 256 pixels). The highest recorded accuracy was at 96.33%. As a result, the presented framework will motivate the finding of a more robust model that automatically and precisely evaluates the condition of sewer pipeline systems using CCTV and encourages the integration of the proposed model in real applications.

1. Introduction

Public infrastructure is the lifeblood of every community, and the underground sewer system is its backbone. Modern underground sewer systems are constructed in the form of a complex pipeline network, and their maintenance is not easy tasks due to the difficulty to detect and diagnose the defects inside the system. A delay in detection and diagnosis of a sewer system can lead to an unexpected functional failure or structural integrity [1], which can cause not only severe damage to the environment but also requires high repairing cost. In the worst scenarios, it can even lead to human casualties. The most critical steps to avoid sewer pipe deterioration is regular inspection of underground pipelines. However, rehabilitation and maintenance of an aging sewer require considerable budget and time-consuming planning process [2].

Presently, in sewer inspection and maintenance, robots are usually deployed to record CCTV videos that can be used to assess the structural condition of sewer pipes later. CCTV systems are widely used due to the unsanitary environment, complex surveillance circumstances, and high pressure under sewer pipes. Moreover, they are the most prevalent and cost-effective methods to assess a sewer system [3]. There is a massive competition between giant industrial robotics companies to develop a new generation of robot that is cheaper, smarter, and more efficient. In Korea, Electronics and Telecommunication Research Institute (ETRI) has currently developed an advanced CCTV inspection technology which was used by UnderGround Safety (UGS) research [34]. The main purpose of the research is to use a utility hole inspection vehicle which can reach up to 50 ft deep in sewerage lines with a diameter of over 600 mm to inspect their condition. The robot provides a 360-degree field of view and a 3D point cloud for precise utility hole measurements. Furthermore, its head is equipped with a high-resolution camera which is capable of inspecting in-pipe defects precisely.

As evaluation technology continues to develop, an automated defect classification system has become a valuable tool to improve performance and save a considerable amount of money for inspection and assessment processes in long-term; it also enables the development of consistent deterioration models and proactive asset management strategies. This system can evaluate recorded CCTV videos and analyzes the sewer line condition automatically. The automated defect classification system must be integrated with existing CCTV software to support the inspection process by providing real-time feedback and notifying the

* Corresponding author.
E-mail address: hmoon@sejong.edu (H. Moon).

https://doi.org/10.1016/j.autcon.2019.102849
Received 26 June 2018; Received in revised form 22 April 2019; Accepted 25 May 2019
0926-5805/ © 2019 Elsevier B.V. All rights reserved.

operator through the defect indication module to avoid possible misinterpreting or skipping defects due to the operator's exhaustion or inexperience.

In a CCTV video, subtitle information is crucial because it provides in-depth details about the condition of the sewer system so an operator can easily observe and pinpoint the location and severity of a defect. A text extraction system is usually applied to extract subtitle information; it includes three phases. Localization focuses on locating text lines, followed by enhancing the image quality to increase the contrast between text and background. Finally, in the recognition phase, optical character recognition (OCR) engines are used to recognize the text. Although OCR engines work best on scanned documents, they do not provide satisfying results if an image has poor resolution, low contrast, or contains too complicated background.

This study proposes a deep learning framework that supports automated defect classification and location recognition in sewer frames extracted from CCTV inspection videos. We also introduce a dataset originated from CCTV videos; it is generated and evaluated manually. Initially, all frames are extracted from CCTV videos, then images which show sewer defects are collected and labeled to a corresponding class. The proposed system includes three main modules. The first module extracts keyframes from CCTV video, and then the second module is implemented to recognize all frames that contain defects and classifies them into a specific class. Finally, for each of the extracted frame, text detection and recognition module are implemented to recognize subtitle information from a frame, which includes position, date, and time of the inspection.

Applying the proposed model, we aim to answer the following questions by conducting various experiments and use results as a foundation:

1. What is the performance of the deep learning based sewer pipe condition assessment on the collected dataset?
2. Are locations of the defects correctly recognized by text recognition module?
3. Does a report on a specific CCTV video generated by the proposed model match a report generated manually by an operator?

The rest of the paper is divided as follows. Section 1 introduces a problem statement for this research. In Section 2, we thoroughly survey previous sewer pipe condition evaluation approaches. The proposed framework will be explained carefully in Section 3. In Section 4, we describe in detail how the proposed dataset was collected as well as evaluation protocols which were used to evaluate experiments results. In Section 5, various experiments are conducted to test the proposed model on the collected dataset. Based on the results from Section 5, Section 6 provides detailed discussions. Finally, Section 7 summarizes entire research and make some comments on future direction.

2. Related work

With the advent of technology, especially in computer vision, the number of vision-based methods to detect defects inside a sewer has increased rapidly. However, conventional computer vision techniques require many pre-processing steps and relevant features need to be selected manually. For example, Zhang [5] applied morphological operations and threshold on grayscale level images to detect potential defective regions, and then they used a distance histogram based shape descriptor to extract defects features. They successfully removed over 90% of misidentified objects and reserved 90% of defects' length. Yang [1] used wavelet transform and computation of co-occurrence matrices to extract text features. Finally, they applied a neural network approach and proved that it performed better than support vector machine (SVM) and the Bayesian classifier. Su [6] used morphological segmentation based on edge detection (MSED) to assist inspectors to detect pipeline defects in CCTV inspection images. They also applied mathematical morphology-based image segmentation methods, which included opening the top-hat operation (OTHO) and closing bottom-hat operation (CBHO). Most segmented cracks had completeness above 50% by CBHO. The highest completeness was 82.79%. Recently, Phat Huynh [7] proposed a novel 3D inspection system to detect anomalies in sewer pipes using stereo vision coupled with novel image processing algorithms and showed that various types of defects were detected successfully. Sinha [8] and Duran [9] applied artificial neural networks for sewer fault detection frameworks. Duran [9] retrofitted CCTV camera with a laser profiler that passed precise internal measurements to ANN to identify structural faults. Sinha [8] identified significant features within the CCTV footage and applied fuzzy logic to some characteristics, such as shape size and light intensity. Then, fuzzy features were fed to the trained ANN to recognize cracks within CCTV footage. Another research proposed an automatic fault detection method for recorded CCTV videos [10]. The authors calculated a feature descriptor for each video frame before passing it to a machine learning classifier to predict contents of a particular frame. They achieved over 80% detection accuracy on still images. Similarly, an anomaly detection approach for sewer fault detection by using CCTV videos in [11]. Authors used a one-class support vector machine (OCSVM) to train the images of regular pipes and highlighted any abnormalities or faults within a sewer video for further analysis.

In recent years, deep learning has been widely used in various computer vision-based tasks, such as object detection and image classification. Deep learning models are capable of extracting visual features automatically from images so unlike conventional machine learning techniques, they do not require many processing steps. Recently, many researchers have applied deep learning based approaches to detect the defects inside civil infrastructures. Cha [12] used a convolutional neural network model to identify cracks on the roads. The model was trained on 40,000 images with an accuracy of about 98%. Moreover, the trained CNN was combined with a sliding window technique to scan an image with a size larger than 256 × 256. Zhang [13] proposed a quantitative evaluation of road defect detection was implemented using a dataset of 500 3264 × 2448 images and achieved an accuracy of over 90%. Moselhi and Tariq [14] applied a three-layer neural network combined with a back-propagation algorithm to classify four types of sewer defects. The accuracy was at over 98%, and it correctly classified 214 out of 218 cases in the testing dataset. Cheng and Wang [15] developed a deep learning-based approach for pipe defect detection via a faster region-based convolutional neural network model (faster R-CNN). They acquired 3000 images, which were collected from CCTV inspection footages. They used 85% for training and validation process whereas 15% was used for testing. Also, they used mean average precision (mAP), missing rate, and detection speed to evaluate system performance and yielded 83% mAP. Although many sewer defect detection has already used CNN, their proposed frameworks only classified defects inside sewer lines, whereas our proposed system classified sewer defect as well as showed defect location by combining deep learning and computer vision techniques.

There have been various approaches for text detection. Two widely used techniques are sliding window classification and connected component analysis (CCA). The connected component-based method considered text information as a set of distinct connected components based on color similarity or spatial layout [16]. On the other hand, in sliding window classification approaches [17], a classifier is fed with positive windows that contain text, and these windows are further divided into text areas by applying morphological operations. Existing methods do solve specific text detection challenges to some extent, but one model worked well on a particular type of dataset but became ineffective on other types, because the critical issue in text detection is the complex background. For subtitle detection in videos, multi-frame integration is usually used, because it reduces background complexity and increases the detection rate. For example, Guo et al. [18] implemented a multi-frame corner matching to lower the impact of the

background on the text. A usual approach is text detection module is applied on several consecutive frames, and then an MFI is used to verify the detected text areas. On another research, maximally stable extremal regions (MSERs) [19] are used as features to extract text or non-text components and proved that the accuracy was significantly improved compared to original features, and it worked best on an image that had a complex background. Moreover, in [20], MSERs in hue, saturation, and value (HSV) color space was applied, and the results showed that HSV color channel outperformed the original red, green, and blue (RGB) color channel in detecting text pixels candidate. This paper investigated MSERs by extracting edges and connections to refine text components.

Building upon previous approaches, this paper proposes a convolutional neural network (CNN) based system for sewer defects classification and location recognition. The framework was constructed by applying transfer learning [21] and fine-tuning existing CNN architecture (see Section 3.1). The proposed model was trained using a total of 47,072 images that were manually extracted from 6605 sewer CCTV videos. In [22], CNN method was applied to categorize sewer CCTV images into three types of defects (root intrusion, deposits, and cracks). In the proposed system, six different types of sewer defects were investigated. Moreover, the proposed system also included a text detection and a recognition module to analyze subtitles printed on the CCTV inspection footage and showed exact defect location.

3. Proposed framework

Fig. 1 shows an overview of the proposed framework. Frames extracted from sewer videos are divided into two classes (normal and defect), then each defect image is assigned a corresponding label. At the end of this step, a huge sewer defect dataset is constructed. After that, a convolution neural network (CNN) model based on the model developed by Krizhevsky [23] is applied to classify defects in underground sewer pipes. The original AlexNet model was trained to classify ImageNet dataset which contained 1.28 million images belonged to a thousand classes [24] — in our framework, fine-tuning and augmentation processes are applied. Besides, text detection and the recognition modules are implemented to recognize the location of a defect after classifying each frame from a CCTV video.

3.1. Deep convolutional neural network for defect classification

CNNs have emerged as critical hierarchical architectures that are capable of learning abstract features from data automatically. CNN models have proved their effectiveness in a wide variety of applications such as segmentation [25], face recognition [26], speech recognition [27], drug discovery [28], and plant disease detection [29]. A typical CNN model consists of three different neural layers, which are convolutional, pooling, and a fully connected layer. Each of them has a specific role in the model architecture. A layer is made of neurons, and visual cortex inspires the connectivity between these neurons. Each neuron has learnable weights and biases, and it accepts some inputs and performs dot product. The last layer is a fully connected layer which is responsible for computing class probabilities. Neurons in a layer act like edge detectors and react to the various types of edges encountered in an image. The inherent hierarchy in deep networks allows neurons in deeper layers to learn more complex structures, which ultimately result in the remarkable performance of CNNs in recognition tasks. Arevalo et al. [30] and others in [31,32] showed that CNNs trained on huge datasets, such as ImageNet, could act as generic descriptor extractors that have powerful discriminative capabilities. The CNN model used in this research was AlexNet which consists of eight learned layers, five of them are convolutional layers, and the remaining layers are fully connected layers; it was proposed in 2012 ImageNet large-scale visual recognition challenge (ILSVRC-2012) and achieved a remarkable performance compared to other non-deep learning approaches in ILSVRC-2012.

The pre-trained AlexNet model is designed to recognize 1000 categories of natural objects in ImageNet dataset. However, in this research, AlexNet is fine-tuned to extract visual features from sewer images. Fine-tuning works on a principle of transfer learning, where CNN models are created to deal with extensive classification problem

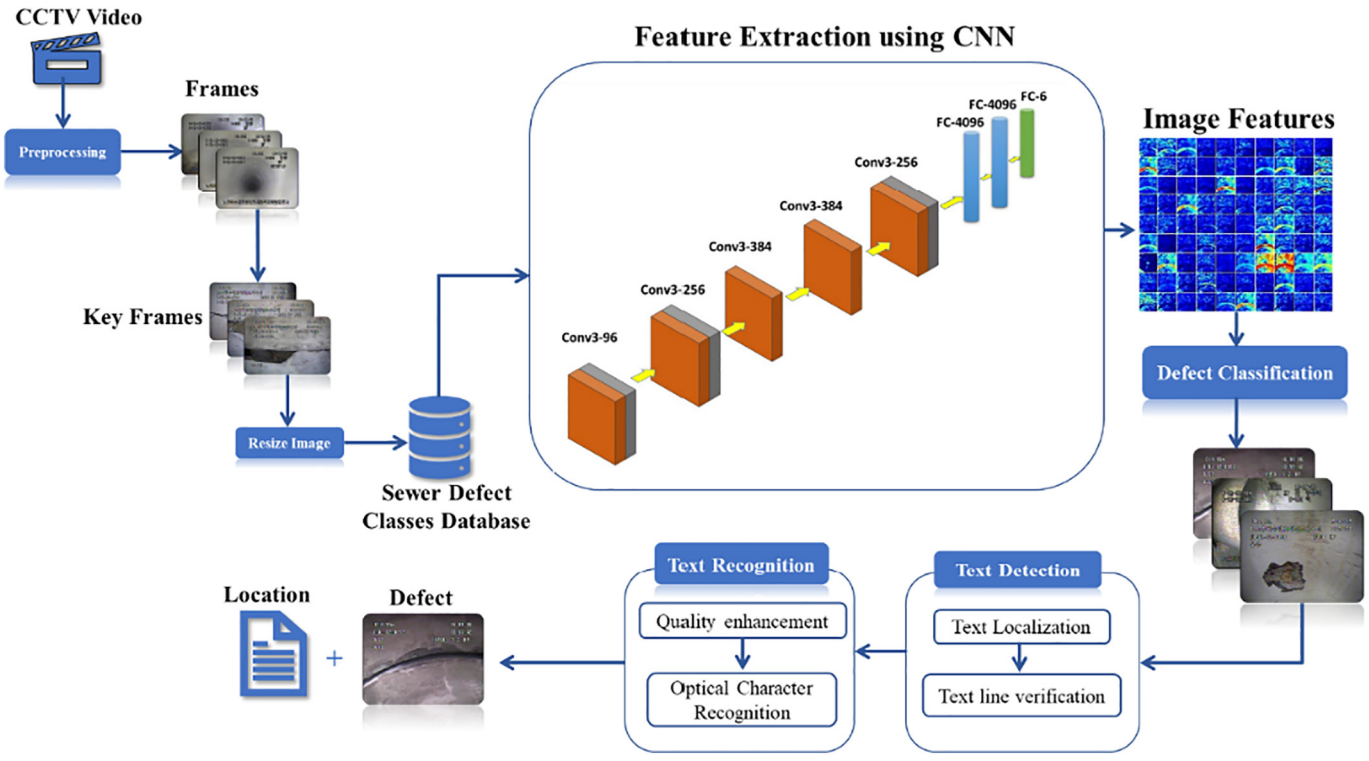

Fig. 1. Overview of the proposed defect classification and defect location recognition framework for sewer line assessment system.

(such as ImageNet classification). These models can be applied in other classification problems as optimized features extractor to minimize the error in the specific domain. In transfer learning, features and parameters from a prior network are transferred to a new network, and the new model can yield high performance and significantly less training time with suitable modifications. Inspired by the transfer learning concept, AlexNet model's parameters are slightly modified. The network is fed with 256 × 256 fixed sized images, and dimension of the last fully connected layer is changed according to the sewer dataset and is set to six output neurons, and each neuron corresponds to one of the six classes from sewer dataset which includes crack longitude, debris, joint faulty, joint open, lateral damage, and surface damage.

Fig. 2 and Table 1 describe a detailed configuration for each layer. The first eight layers of AlexNet architecture, which are Conv1, Pool1, Conv2, Pool2, Conv3, Conv4, Conv5, and Pool3 are dedicated to features extraction. After each convolutional layer, a ReLU (Rectified Linear Units) function is applied at the output of the convolutional layer. There are three fully connected layers (FC-1, FC-2, and FC-3), and the output of the last fully connected layer was reduced from 1000 to 6 neurons.

To train and test the CNN model, an NVIDIA DIGITS 5 toolbox with Caffe framework was used. Experiments are performed on an Ubuntu 14.04 OS that used an Intel® Core i7-5930K processor, four NVIDIA Titan XP 12GB GPUs, four 3072 Cuda cores, and 64GB of DDR4 RAM.

3.2. Text detection and recognition

We used text detection and recognition framework proposed by Dang in [35]. As depicted in Fig. 3, the model consisted of four Section 1) In multi-frame integration section, for each second, 30 continuous frames were extracted from the input video. After that, frame averaging was adopted to enhance text edge and reduce background complexity. 2) Image preprocessing. 3) Text detection included two steps, which were text localization (find the text lines) and text line verification (all detected false alarm lines are removed). 4) Text recognition consisted of two steps, which were text quality enhancement and training with Tesseract OCR.

Although CCTV videos were recorded under various environments, captions always appeared in every frame at a fixed position whereas background changed continuously as the robot moved forward in a sewer line. The robot recorded a video at 30 Fps (frames per second). Thus, a multi-frame integration (MFI) method was applied to a patch of 30 continuous frames. Moreover, within these 30 frames, subtitle information was guaranteed to be the same. Frame averaging technique used in this study is multi-frame average.

For a frame cluster C_i (from frame i to frame $i + 29$), the output image is generated as follows:

$$AverageImage_i(x,y) = \underset{j \in C_i}{avg}\left(p_j(x,y)\right) \quad (1)$$

let $p_j(x,y)$ indicates the pixel value of frame j at position $(x,y.)$

Fig. 4 shows two examples of a multi-frame averaging method. Fig. 4(a) describes the frames before applying the multi-frame integration technique; the background of the image is quite complex and contains many edges, which significantly lower text detector performance. However, after applying multi-frame averaging, the background's complexity was vastly reduced as presented in Fig. 4(b). Finally, Fig. 5 shows the result of text detection after applying the method proposed by [35].

In the text recognition module, Tesseract OCR [36] was implemented, which is an open-source OCR engine based on long short-term memory algorithm that was developed at HP and has recently taken over by Google. It supports the training of text recognition for various languages. The reason we used Tesseract OCR instead of other text recognition engines was due to its impressive performances on various research, such as in [19,35].

4. Proposed dataset and evaluation protocols

4.1. Proposed dataset description

4.1.1. Dataset acquisition

In this research, a total of 6605 CCTV videos that can be to access sewer pipes lines' condition are used; it is provided by Korea Institute of Civil Engineering and Building Technology.[1] These videos were taken by a commercially available Robo Cam 6 (Tap Electronics Ind. Co., Ltd); it is equipped with a 1/3-in. SONY Exmor CMOS camera module with camera capability to capture 360° continuous rotations, and 240° side views up/down tilt. It also uses powerful halogen lamps to capture the images/videos in various lighting conditions. The duration of each video is from 1 to 15 min and subtitles information printed on each video contains essential information for further inspection. Table 2 shows eight types of subtitle information on CCTV videos.

4.1.2. Data augmentation

The effectiveness of deep CNN models is known to depend on the availability of large training data. As a result, data augmentation is a useful technique to expand training data. A significant attribute of the data augmentation is that predefined classes remain unchanged after applying those augmentation techniques. It has been shown that data augmentation can reduce overfitting on a model and increase the amount of training data. There are various augmentation methods, such as transformation, deformation, flipping, rotation, and translation. Wang et al. [33] applied generative adversarial networks (GANs) and standard transformations to create a large dataset. They used a horizontal flip to increase the training data. However, the rotation was not adopted because types of defects were sensitive to rotational flip. For example, if the rotation is applied to the joint faulty class, the joint faulty class became crack longitude. It will be confusing and may reduce the performance of the model. Based on Wang et al. [33] research, we applied the horizontal flip to the proposed dataset. Furthermore, text detection and recognition modules are not affected by flipped images because the original dataset was used in this module.

4.1.3. Dataset description

In the proposed dataset, six types of sewer defects were investigated (Fig. 6). The total number of manually validated defected images before applying augmentation was 24,137. Then, the number of images in the dataset increased to 48,274 after data augmentation was used. Furthermore, the dataset was divided into two separate parts, one for training and the other for testing. In the training part, 97% of the entire dataset, which equals to 47,072 images, were applied for training and validation purpose (75% out of 47,072 images were used as training, and 25% were used as validation). The remaining 3% of the dataset was used in the testing part (1202 images). Fig. 6. depicts 6 types of sewer defects extracted from CCTV videos, and Table 3 describes the details of the sewer defect dataset.

Because frames extracted from different videos varied in size, so all images were resized to a fixed size of (256 × 256 pixels). The reason for choosing a relatively small size is that the CNN models are usually trained on an image with the resolution from 128 × 128 to 256 × 256 pixels. Besides, although higher resolution images provide more precise information compared to lower resolution images, they require high computational power and a significant amount of processing time.

4.2. Evaluation protocols

4.2.1. Sewer defects classification

This section describes the evaluation protocols that were used to

[1] [2] https://www.kict.re.kr/eng.

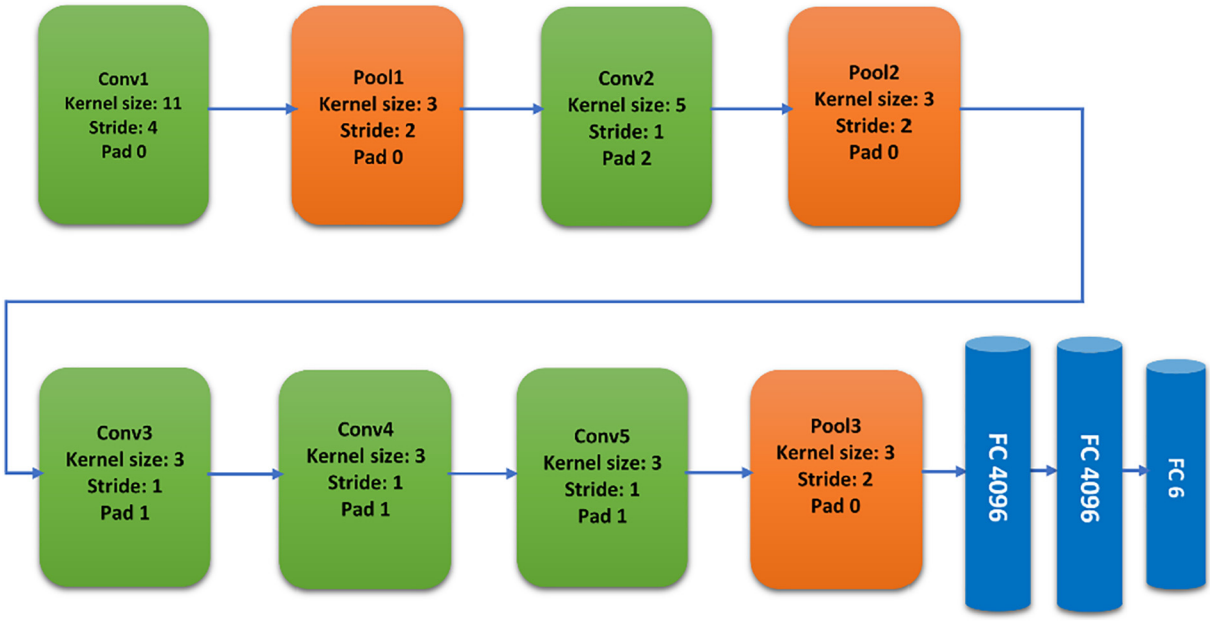

Fig. 2. CNN architecture for the proposed sewer defect classification framework.

Table 1
Detailed configurations of the proposed CNN model.

Configuration	Conv2	Pool1	Conv2	Pool2	Conv3	Conv4	Conv5	Pool3
Input map	3	96	96	256	256	384	384	256
Input	256 × 256	55 × 55	27 × 27	27 × 27	13 × 13	13 × 13	13 × 13	13 × 13
Output map	96	96	256	256	384	384	256	256
Filters	11 × 11	3 × 3	5 × 5	3 × 3	3 × 3	3 × 3	3 × 3	3 × 3
Stride	4 × 4	2 × 2	1 × 1	2 × 2	1 × 1	1 × 1	1 × 1	2 × 2
Zero padding	0	0	2	0	1	1	1	0

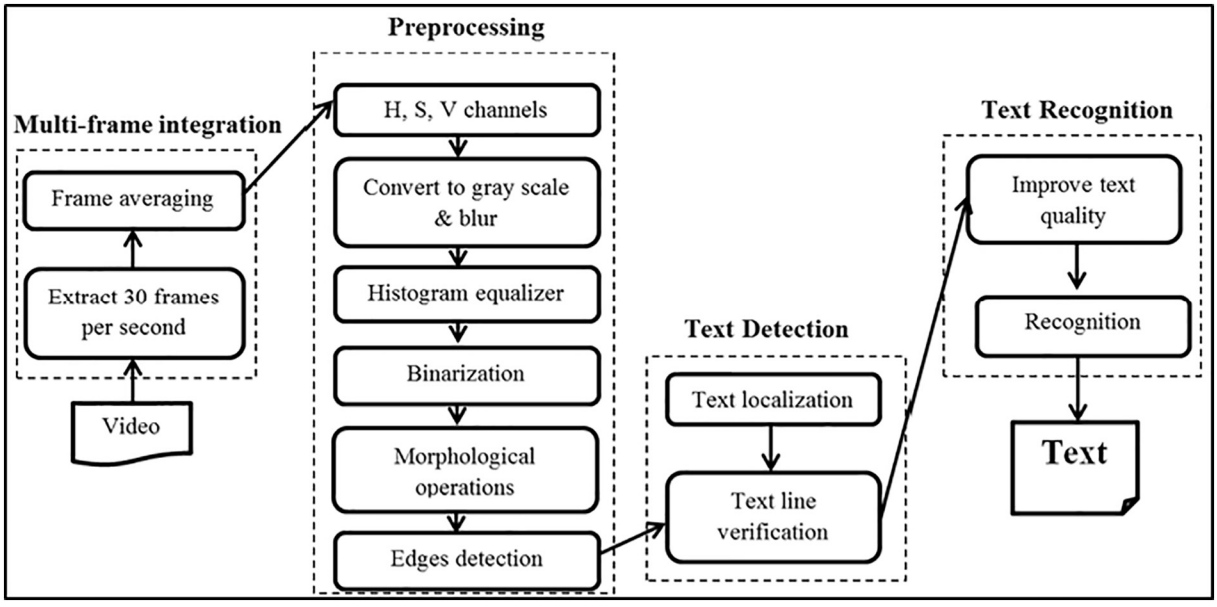

Fig. 3. The system architecture for the sewer text detection and recognition [35].

evaluate sewer defects classification as well as text recognition. The performance of the defect classification framework was evaluated using six videos from the sewer CCTV videos. Results generated from the model were compared with manually generated assessment reports to examine the proposed framework's effectiveness. These reports were created by UnderGround Safety (UGS) [4] research at Electronic and Telecommunication Research Institute (ETRI) South Korea. An operator inspects each video manually and checks whether defects appear. If the video includes any defects, then the operator classifies those defects according to the type of sewer defect.

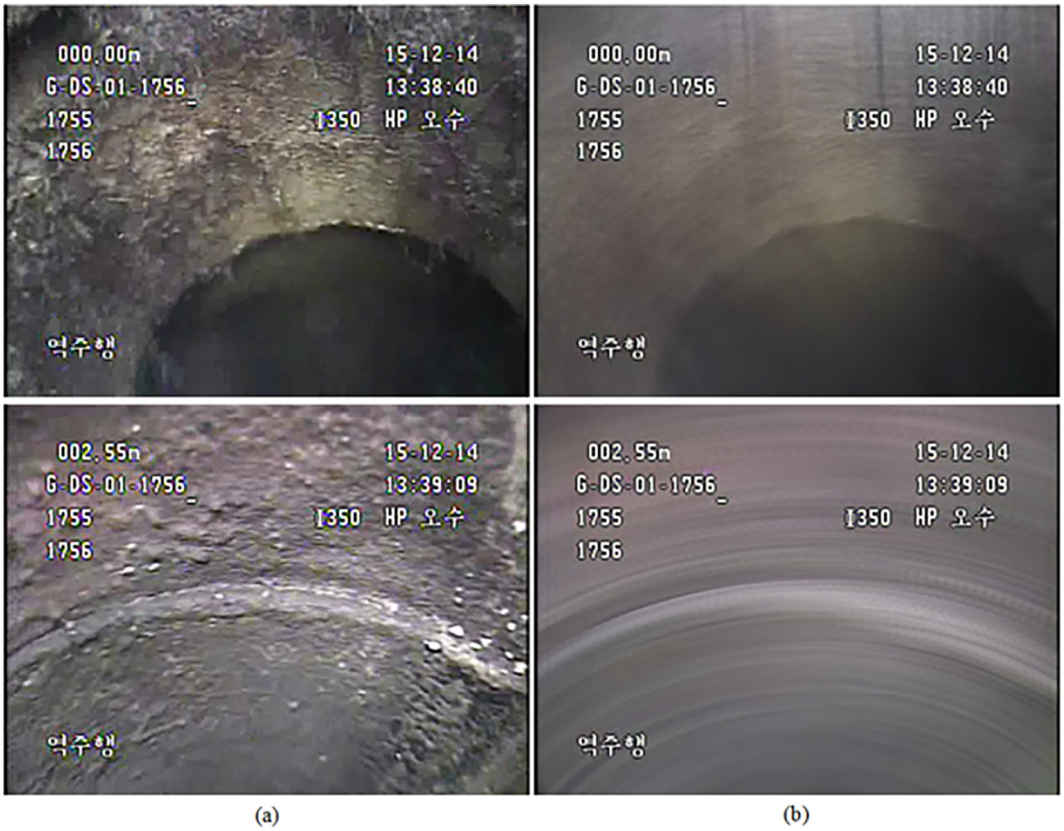

Fig. 4. Sample images before and after applying the multi-frame averaging (a) Without the multi-frame averaging and (b) With the multi-frame averaging [35].

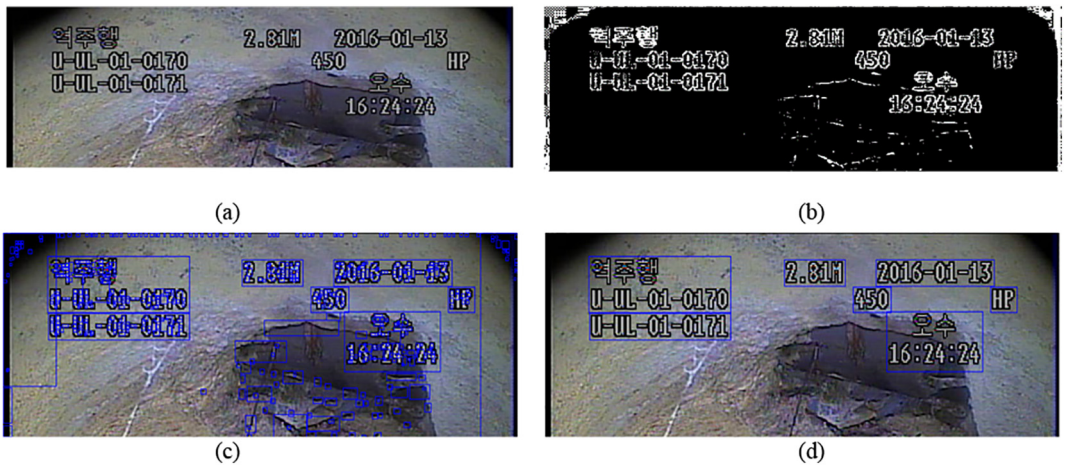

Fig. 5. Text line verification using saturation channel. (a) Original image, (b) The saturation channel, (c) Text lines detection results (blue bounding box), and (d) Results after applying text lines verification (blue bounding box) [35]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Subtitle information in the text recognition module.

Subtitle information	Contents
Driving distance	Distance from the starting point
Pipe number	Pipe unique number
Survey date	Date of an investigation
Survey time	Survey time
Circumference	Size of sewer line
Type	Type of sewer lines
Start/end location	Start and end location of the exploration
Driving direction	Backward or forward

4.2.2. Text detection and text recognition

The evaluation protocol used in text detection and recognition was recommended by Wolf et al. in [34]. The approach shows object level precision and recall using detection quality restraints. Both the amount and the feature of the detected bounding boxes were calculated. The assessment was calculated by precision, recall, and F-measure as follows:

$$Precision = \frac{\sum_i^N \sum_j^{|D^i|} M_D(D_j^i, G^i)}{\sum_i^N |D^i|} \quad (2)$$

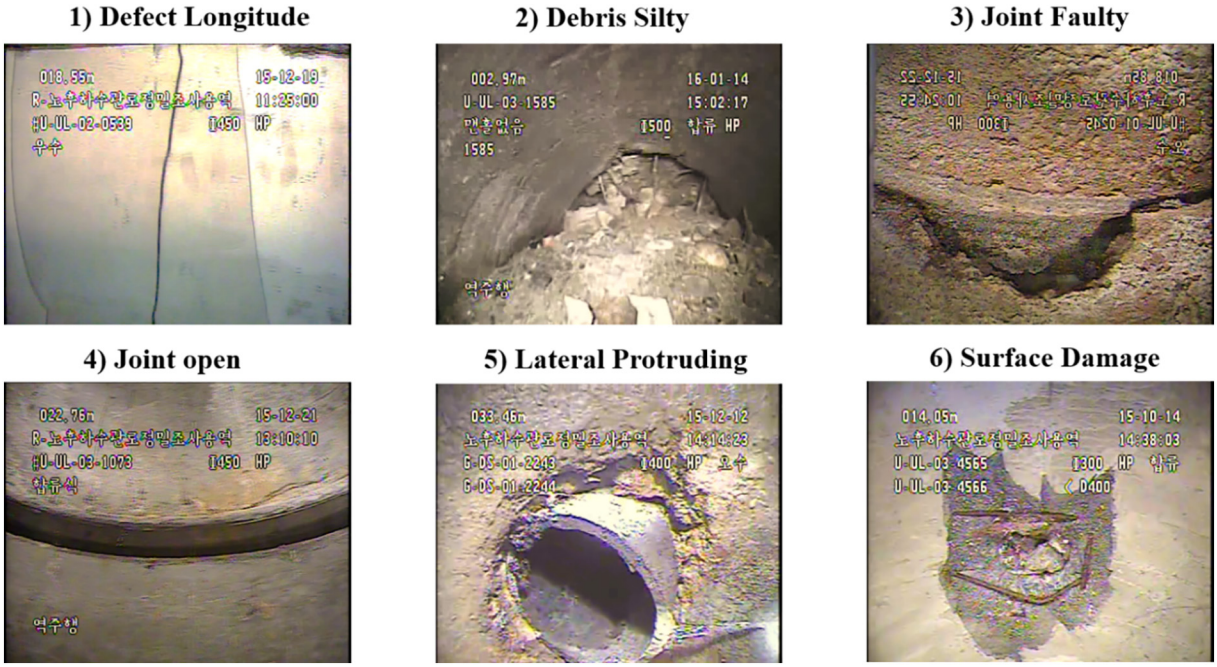

Fig. 6. Six types of sewer defects extracted from CCTV videos.

Table 3
Detailed description for the sewer defect dataset.

Defects class	No. images	After augmentation	Training images	Testing images
Longitudinal defect	2265	4530	4530	200
Debris silty	3882	7764	7764	200
Joint faulty	3801	7602	7602	200
Joint open	6146	12,292	12,292	200
Lateral	4767	9534	9534	200
Surface damage	3276	6552	6552	200

$$Recall = \frac{\sum_i^N \sum_j^{|G^i|} M_G(G_j^i, D^i)}{\sum_i^N |G^i|} \quad (3)$$

$$F_{measure} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

where N is dataset size. $|D^i|$ and $|G^i|$ are the amount of detected and ground truth rectangles in image i-th. $M_D(D_j^i, G^i)$ and $M_G(G_j^i, D^i)$ are the matching scores for the detected rectangle D_j and ground truth rectangle G_j. Two rectangles are considered equal if their intersection proportion is higher than a fixed threshold, which manages matching quality. The threshold for one to many matching was set to 0.8 for the simple background dataset and 0.6 for the complex background dataset.

The number of correctly recognized letters measured the *word* recognition performance *(WRA)*, and it is defined as:

$$WRA = \frac{|C|}{|T|} \quad (5)$$

where C indicates the amount of correctly recognized letters, and T is the number of ground truth letters.

5. Experimental results

5.1. Sewer defect classification

In previous research on defect classification, learning features were selected and extracted manually. However, CNN models learn to extract features automatically by updating the weights of receptive fields [34]. In this paper, 75% of the dataset was randomly selected for training, and 25% of the dataset was used to validate the training process and learning rate. The learning rate can be defined as the optimization and minimization of the loss function of the network. Moreover, it is the most crucial hyper-parameter for tuning the networks, and it determines how fast weights (in the case of a neural network) or the coefficients (in the case of linear regression or logistic regression) change [37–39]. In the proposed model, initially, the learning rate was set to 0.01, and it was gradually reduced to 0.001 according to the error rate of the validation set.

The accuracy of training and validation phases increased significantly to over 80%, and the corresponding loss of training and validation decreased dramatically to below 10% after the first ten epochs. Then the accuracy increases gradually before stopping at over 96% while the loss decreases constantly to 10%. The highest accuracies achieved in the training and validation process were 96.50% at the 25th epoch and 96.60% at the 30th epoch, respectively. Fig. 7 summarizes the training and validation accuracies. In this study, the CNN model was trained with 30 epochs, and the total training time lasted 1 h 35 min.

Fig. 8 describes the class activation map of six common sewer defects, which are lateral, joint open, joint faulty, debris, surface damage, and defect longitude. For each defect, the corresponding class activation map shows that the defect classification framework correctly learned the defects features. Besides, Table 4 presents a confusion matrix for the testing set, and it was computed to assess the ability of the proposed CNN model in classifying different defects. Experimental results suggested that the CNN model correctly recognized all six classes with the highest accuracy at 99.5% on both debris silty and surface damage. However, there is a low accuracy at 85% on defects longitude class.

In addition to the previous experiment, our proposed system was further applied to a total of six videos from the sewer CCTV video database. Results were then compared with assessment reports. Table 5 shows the comparison between results from our model and results collected from assessment reports. Overall, the number of defects identified by our model is equal to the number of defects showed in reports. Especially, in video number 5, precisely 13 defects similar to

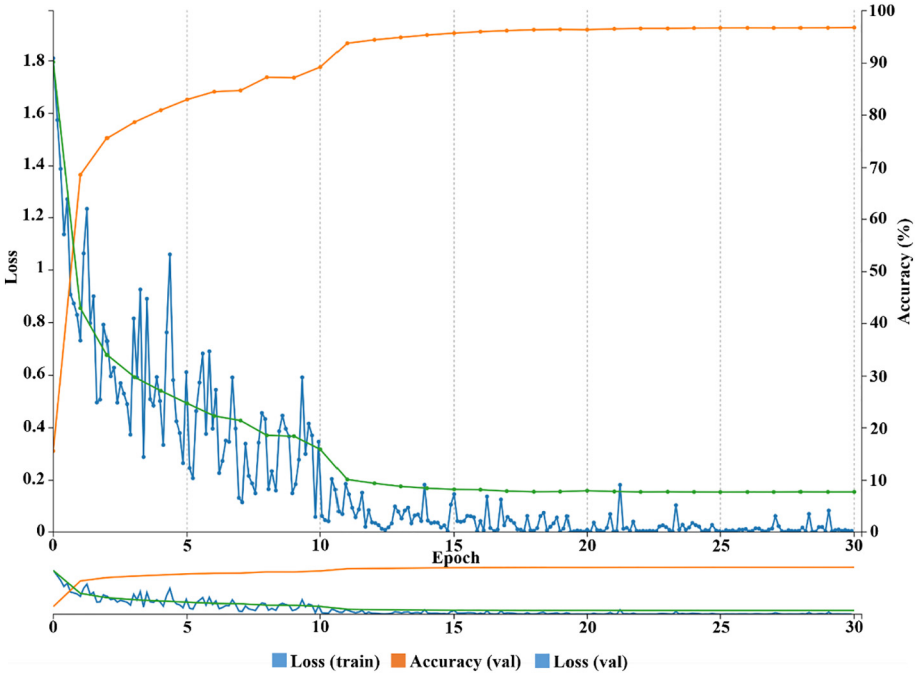

Fig. 7. Training and validation accuracies for each epoch.

the report are detected by the proposed model.

5.2. Text detection on CCTV's recorded videos

Detection and recognition module was used to recognize subtitle information from ten videos. Although the robot started recording on the ground for some time before it was put into the sewer line to check the recording quality, we only need to extract frames after the robot was in the sewer line. Thus, the frames, which contained the "start the inspection" subtitle in Korean, were searched. When this information was detected in a specific frame, all frames (after that frame) were extracted. On the contrary, all video frames will be extracted.

All the videos were recorded at 30 frames/s on the sewer system at different locations. Information regarding the length of the video, the total extracted frames, the number of frames after detecting the "start of inspection" Korean text, and the number of frames after applying multi-frame average are described in Table 6.

Subtitle information is described in Table 7. In each video, only the travel distance changed as the robot was moving forward, while other information was similar throughout the video. We also selected two videos (Video ID 7 and 8) as shown in Fig. 9, which had a different font compared to the rest of the videos in the dataset to check whether Tesseract OCR can recognize text information.

Text detection module was used on each video. Ground truth labels were manually created, and then they were used as the ground truth to compare with detected bounding boxes. The model detected 41,058 bounding boxes out of 46,328 bounding boxes with false alarms of 5270 text boxes. The result proved that using the multi-frame average technique significantly reduced wrongly detected boxes and simultaneously increased the accuracy. Enhancement of background quality increased the quality of the low-resolution text, but at the same time blurring the background. Also, the detection module can detect text boxes in a video that has a slightly different font and format.

5.3. Blurred images analysis

In this experiment, we evaluated the performance of the CNN model under extreme illumination conditions by applying an artificial blur on testing images as shown in Fig. 10. Table 8 describes the observed results before re-training the model with blurred images. It was

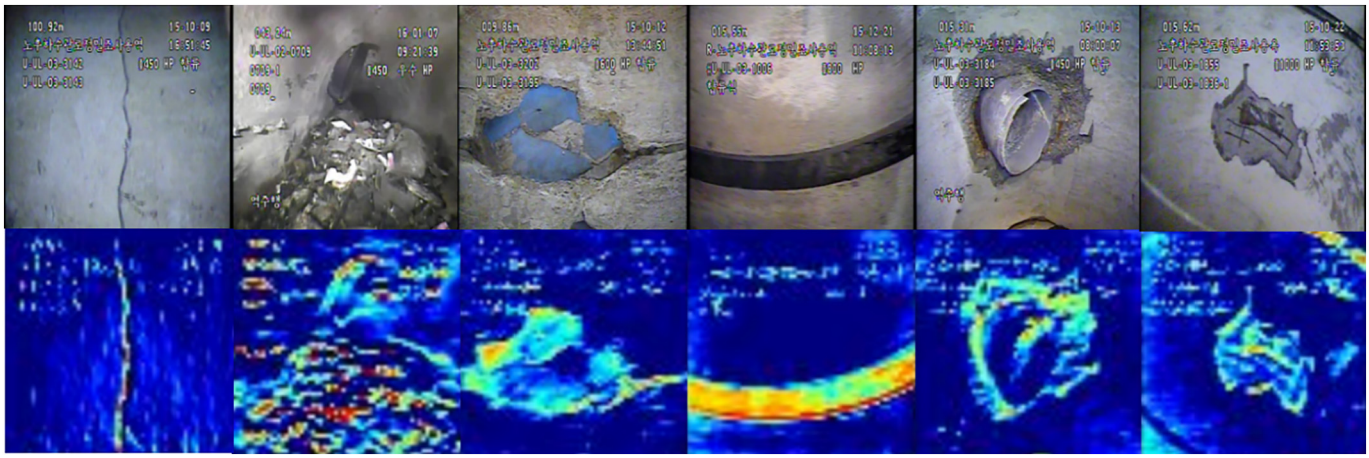

Fig. 8. Sewer defects and their corresponding visualization.

Table 4
Confusion matrix on the sewer defect dataset (testing) using the proposed defect classification system.

Class	Longitudinal defect	Debris silty	Joint faulty	Joint open	Lateral	Surface damage	Accuracy
Longitudinal defects	170	0	0	25	0	5	85%
Debris silty	0	199	0	0	1	0	99.5%
Joint faulty	2	0	198	0	0	0	99%
Joint open	0	0	7	193	0	0	96.5%
Lateral	0	3	0	0	197	0	98.5%
Surface damage	0	0	1	0	0	199	99.5%
Average accuracy							96.33%

Table 5
A comparison between results from proposed model and manually generated assessment report.

#	Video name	No. defects in report	No. defects from the proposed model
1	G-DS-01-0707~0708	1	1
2	G-DS-02-1896	4	3
3	G-DS-02-1899	4	4
4	G-DS-02-2043	4	4
5	G-DS-01-1531	13	13
6	G-DS-01-2180	3	3

Table 6
Number of frames before and after applying multi-frame integration.

#	Video length (Seconds)	Total extracted frames	Frames after detecting "start of inspection."	Total frames after applying MFI
1	682	20,483	18,234	607
2	675	20,245	16,681	556
3	418	12,532	11,125	370
4	303	9108	8312	277
5	526	15,795	14,963	498
6	680	20,404	16,325	544
7	563	16,889	14,823	494
8	597	17,910	15,121	504
9	639	19,164	19,164	638
10	1304	39,099	39,099	1303

Table 7
Detailed description of CCTV video subtitles.

#	Inspection date	Sewer pipe ID	Diameter	Travel distance
1	15-11-27	GDS012169~GDS012180	I400	000 m– > 052 m
2	15-12-07	GDS011531	I700	000 m– > 047 m
3	15-12-16	GDS010622	I300	000 m– > 027 m
4	15-11-27	GDS010707~GDS010708	I300	000 m– > 041 m
5	15-12-24	GDS012126~GDS012127	I300	0 m– > 41 m
6	15-12-28	GDS012129~GDS012130	I600	0 m– > 63 m
7	15-11-28	GDS010283~GDS010284	I300	0 m– > 48 m
8	15-11-30	GDS010333~GDS010290	I300	0 m– > 42 m
9	15-11-25	GDS011365	I400	0 m– > 47 m
10	15-11-27	GDS011313	I300	0 m– > 58 m

noticeable that defects longitude and joint open classes were affected by the effect of blurred images because they showed poor performance at 51% and 57%, respectively.

However, after adding blurred images to the training dataset and re-training the CNN model, we obtained a significant improvement in terms of accuracy as represented in Table 9. The accuracy for defects longitude class increased from 51% to 86.5%, while the accuracy for joint open class improved from 57% to 99%. This experiment showed that the model's ability to deal with extreme illumination conditions had improved remarkably after the model was trained on blurred images.

5.4. Defects classification and defect location recognition

In this experiment, the performance of the system was assessed by integrating the defect classification and the text recognition module. A sewer video was randomly selected from sewer CCTV videos dataset, then defect classification and text recognition modules were applied to evaluate the interpretation of the system. Finally, the results of the proposed system were compared with the manually generated sewer report as depicted in Fig. 11. Table 10 demonstrates the comparison results of the proposed system and manually generated report results. The distance in the report (xxx.xx m) was a little different from the automatic recognition (xxx m) because we only considered the first three digits of the distance that already indicated the exact location of the defect.

5.5. Comparative analysis of defects classification

Compare to Moselhi [14] results; the proposed method showed a remarkable performance on defect classification. Moreover, Moselhi experimented three sewer defects, which were crack, joint displacement, and spalling, whereas the proposed method could classify six types of sewer defects as well as recognize their locations. Table 11 shows the comparative analysis of the proposed method with Moselhi defect classification results.

6. Discussion

As discussed in the introduction section, three key research questions need to be answered based on the experiments. The first question was about the performance of the proposed model on the collected sewer dataset. The results showed that our model achieved a state-of-the-art performance at 96.3%. The second question and the third question asked about the performance of defects location recognition. As shown in Section 5.4, the results obtained when compared to the manually generated report with the results generated from our model proved that our model performed well on the defects location recognition.

Through various experiments, we proved that our proposed model was effective in detecting sewer defects. We also solved the location recognition that previous research failed to solve. It has a great possibility to reduce the labor cost associated with manually reviewing the CCTV video. As a result, it reduces processing time and labor cost.

7. Conclusion

This paper presented a framework for automated sewer defect classification and recognition of defect location in CCTV inspection videos based on deep learning. Conventional image processing techniques relied heavily on handcrafted feature extraction and morphological methods that did not provide satisfying results when CCTV videos have a complex background and illumination conditions. The proposed system overcomes these challenges by exploiting the deep convolutional neural network approach. In previous studies, most of the research has been done using defects detection and classification, whereas

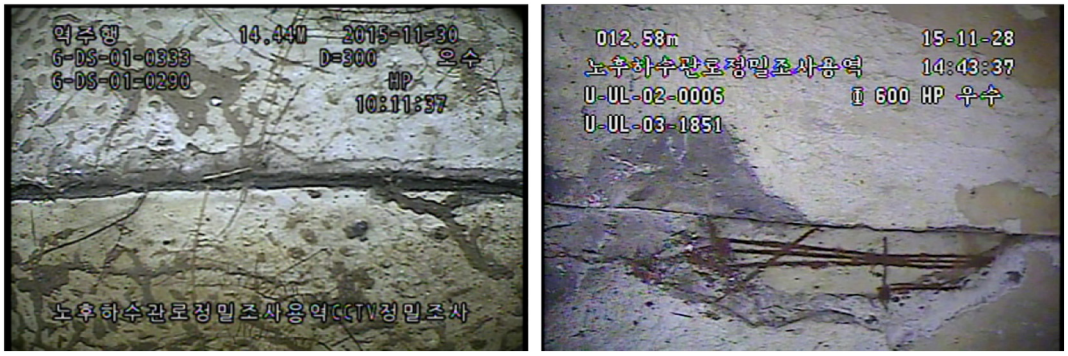

Fig. 9. Videos that use a different font and format compared to other videos in the dataset [35].

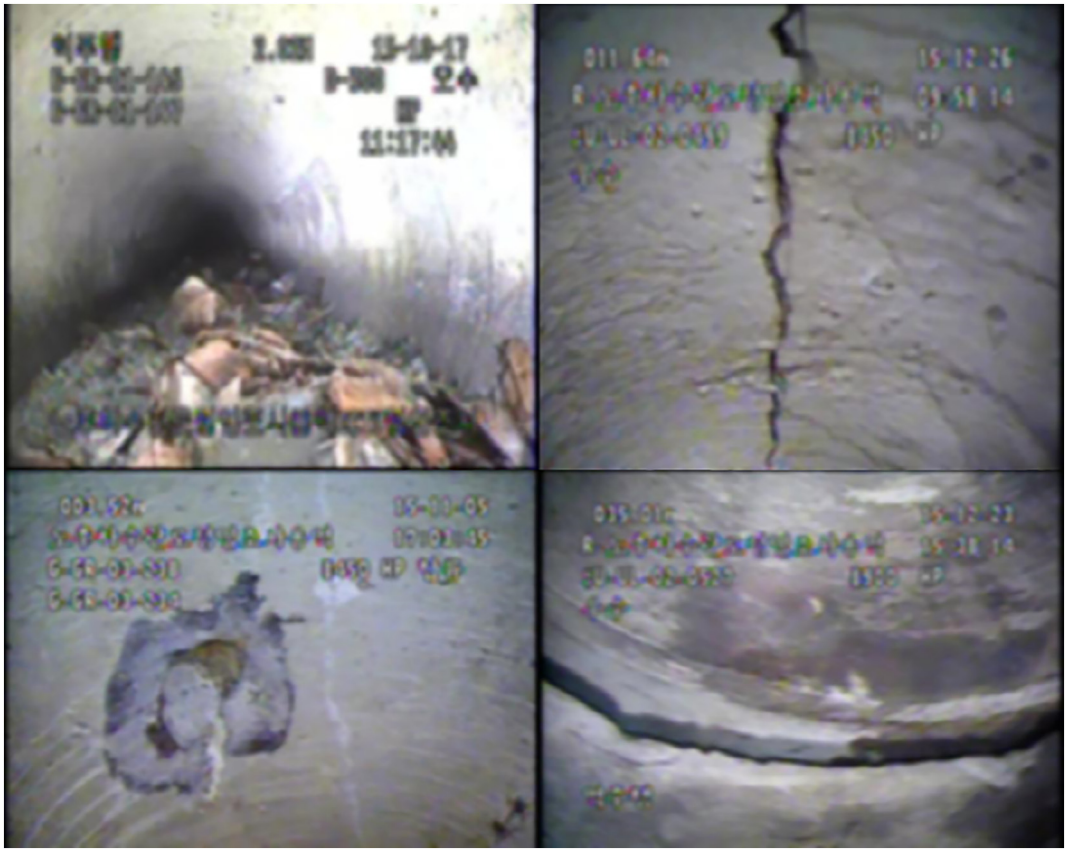

Fig. 10. Blurred image samples resulting from extreme lighting environments.

Table 8
Sewer defect classification results before re-training the model on blurred images.

Class	Defect longitude	Debris silty	Joint faulty	Joint open	Lateral	Surface damage	Accuracy
Defects longitude	102	1	15	11	0	71	51%
Debris silty	0	187	1	8	4	0	93.5%
Joint faulty	0	0	196	2	0	2	98%
Joint open	2	0	78	114	4	2	57%
Lateral	0	0	0	0	198	2	99%
Surface damage	0	0	1	6	15	178	89%
Average accuracy							81.2%

the proposed method not only classifies the sewer defects but also recognizes their location by employing text detection and recognition modules.

With defects classification, a total of 48,274 images that contained defects extracted from different sewer CCTV videos, which contained 48,274 images. Testing results showed consistent performance, even though testing images had different illumination conditions and background noise. The highest accuracy recorded on trained CNN network was 96.33%. Moreover, in text detection and recognition modules, which is the combination of multi-frame integration, various processing

Table 9
Sewer defect classification results after re-training the model on blurred images.

Class	Defect longitude	Debris silty	Joint faulty	Joint open	Lateral	Surface damage	Accuracy
Defects longitude	173	0	2	22	0	3	86.5%
Debris silty	4	188	0	7	1	0	94%
Joint faulty	0	0	198	2	0	0	99%
Joint open	0	0	1	199	0	0	99%
Lateral	0	0	0	0	200	0	100%
Surface damage	0	0	0	0	0	200	100%
Average accuracy							96.58%

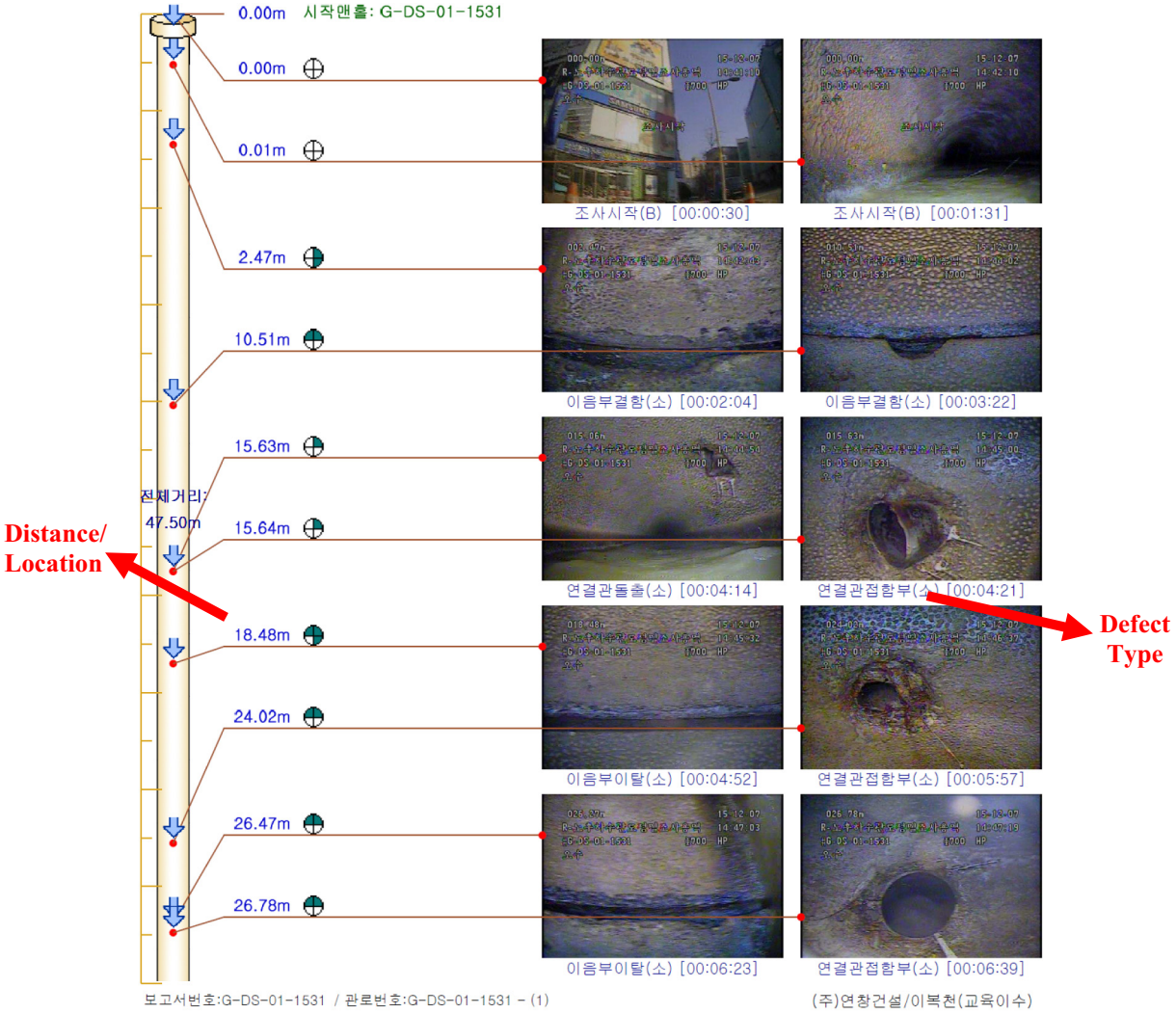

Fig. 11. An example of a manually generated sewer inspection report.

Table 10
Defect classification and location recognition from the model in comparison with a manually generated report.

Classified defects (Manual)	Classified defects (Automatic)	Defect location (Manual)	Recognized location (Automatic)
Joint faulty	Joint faulty	2.47 m	2 m
Joint faulty	Joint faulty	10.51 m	10 m
Debris	Debris	15.63 m	15 m
Joint open	Joint open	18.48 m	18 m
Lateral	Lateral	24.02 m	24 m
Joint open	Joint open	26.47 m	26 m
Lateral	Lateral	26.78 m	26 m

Table 11
Comparative analysis of the proposed defect classification method with other research.

Class	Moselhi DB		Our DB	
	Moselhi et al. [14]	Proposed	Moselhi et al. [14]	Proposed model
Crack	0.9590	0.986	0.806	0.85
Joint displacement	0.9617	0.995	0.896	0.965
Debris	N/A	N/A	N/A	0.995
Lateral	N/A	N/A	N/A	0.985
Surface damage	N/A	N/A	N/A	0.995
Joint faulty	N/A	N/A	N/A	0.99

steps, and MSERs to extract the text edges made the method a truly robust one. The low false alarms rate will ensure the method provides accurate information for the real sewer analyst application.

Furthermore, this study was validated with the Korean language by performing the detection and recognition of Korean subtitles using a multi-scale template matching method. The detection module obtained single-line text instead of a text region that contained multiple text lines, which benefits the recognition module. Although this study was designed mainly for detecting and recognizing text in sewer CCTV's videos, it worked properly for most of the complex background videos. Besides, the proposed system was developed to classify sewer defects. However, if more than two types of defects appeared in an image, the system can classify only single defects, which have the highest probability. In the future, the proposed work can be extended to real-time sewer defects classification to classify multiple defects at the same time. Moreover, the system will be able to cope with live video streaming instead of recordings, which can assist the operators during the inspection process and overcome issues related to operator fatigue and inadequate training.

Acknowledgment

This work was supported by the National Research Council of Science & Technology (NST) grant by the Korea government (MSIP) (No. CRC-14-02-ETRI).

References

[1] Tafuri, Anthony N., and Ariamalar Selvakumar. "Wastewater collection system infrastructure research needs in the USA." *UrbanWater* **4** (1) 2002, 21–29. https://doi.org/10.1016/S1462-0758(01)00070-X.

[2] Ming-Der Yang, Tung-Ching Su, Automated diagnosis of sewer pipe defects based on machine learning approaches, Expert Syst. Appl. 35 (3) (2008) 1327–1337 https://doi.org/10.1016/j.eswa.2007.08.013.

[3] Dae-Hyun Koo, Samuel T. Ariaratnam, Innovative method for assessment of underground sewer pipe condition, Autom. Constr. 15 (4) (2006) 479–488 https://doi.org/10.1016/j.autcon.2005.06.007.

[4] Electronics and Telecommunications Research Institute website, Available at https://www.etri.re.kr/eng/sub6/sub6_01020101.etri?departCode=61 , Accessed date: 7 January 2019.

[5] Wenyu Zhang, Zhenjiang Zhang, Dapeng Qi, Yun Liu, Automatic crack detection and classification method for subway tunnel safety monitoring, Sensors 14 (10) (2014) 19307–19328 https://doi.org/10.3390/s141019307.

[6] Tung-Ching Su, Ming-Der Yang, Application of morphological segmentation to leaking defect detection in sewer pipelines, Sensors 14 (5) (2014) 8686–8704 https://doi.org/10.3390/s140508686.

[7] Phat Huynh, Robert Ross, Andrew Martchenko, John Devlin, 3D anomaly inspection system for sewer pipes using stereo vision and novel image processing, 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA), IEEE, 2016, pp. 988–993, , https://doi.org/10.1109/ICIEA.2016.7603726.

[8] Sunil K. Sinha, Paul W. Fieguth, Neuro-fuzzy network for the classification of buried pipe defects, Autom. Constr. 15 (1) (2006) 73–83 https://doi.org/10.1016/j.autcon.2005.02.005.

[9] Olga Duran, Kaspar Althoefer, Lakmal D. Seneviratne, Automated pipe defect detection and categorization using camera/laser-based profiler and artificial neural network, IEEE Trans. Autom. Sci. Eng. 4 (1) (2007) 118–126 https://doi.org/10.1109/TASE.2006.873225.

[10] Joshua Myrans, Richard Everson, Zoran Kapelan, Automated detection of faults in sewers using CCTV image sequences, Autom. Constr. 95 (2018) 64–71 https://doi.org/10.1016/j.autcon.2018.08.005.

[11] Joshua Myrans, Zoran Kapelan, Richard Everson, Using automatic anomaly detection to identify faults in sewers, WDSA/CCWI Joint Conference Proceedings, vol. 1, 2018 https://ojs.library.queensu.ca/index.php/wdsa-ccw/article/view/12030.

[12] Young-Jin Cha, Wooram Choi, Oral Büyüköztürk, Deep learning-based crack damage detection using convolutional neural networks, Computer-Aided Civil and Infrastructure Engineering 32 (5) (2017) 361–378 https://doi.org/10.1111/mice.12263.

[13] Lei Zhang, Fan Yang, Yimin Daniel Zhang, Ying Julie Zhu, Road crack detection using deep convolutional neural network, 2016 IEEE International Conference on Image Processing (ICIP), IEEE, 2016, pp. 3708–3712, , https://doi.org/10.1109/ICIP.2016.7533052.

[14] Moselhi, Osama, and Tariq Shehab-Eldeen. "Classification of defects in sewer pipes using neural networks." J. Infrastruct. Syst. 6, no. 3 (2000): 97–104. https://ascelibrary.org/doi/abs/10.1061/(ASCE)1076-0342(2000)6:3(97).

[15] Jack C.P. Cheng, Mingzhu Wang, Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques, Autom. Constr. 95 (2018) 155–171 https://www.sciencedirect.com/science/article/pii/S0926580518303273.

[16] Hyung Il Koo, Duck Hoon Kim, Scene text detection via connected component clustering and nontext filtering, IEEE Trans. Image Process. 22 (6) (2013) 2296–2305, https://doi.org/10.1109/TIP.2013.2249082.

[17] Yang Zheng, Qing Li, Jie Liu, Heping Liu, Gen Li, Shuwu Zhang, A cascaded method for text detection in natural scene images, Neurocomputing 238 (2017) 307–315 https://doi.org/10.1016/j.neucom.2017.01.066.

[18] Zhe Guo, Yuan Li, Yi Wang, Shu Liu, Tao Lei, Yangyu Fan, A method of effective text extraction for complex video scene, Math. Probl. Eng. 2016 (2016), https://doi.org/10.1155/2016/2187647.

[19] Michael Opitz, Markus Diem, Stefan Fiel, Florian Kleber, Robert Sablatnig, End-to-end text recognition using local ternary patterns, MSER and deep convolutional nets, 2014 11th IAPR International Workshop on Document Analysis Systems, IEEE, 2014, pp. 186–190, , https://doi.org/10.1109/DAS.2014.29.

[20] Houssem Turki, Mohamed Ben Halima, Adel M. Alimi, A hybrid method of natural scene text detection using MSERs masks in HSV space color, Ninth International Conference on Machine Vision (ICMV 2016), vol. 10341, International Society for Optics and Photonics, 2017, p. 1034111, , https://doi.org/10.1117/12.2268993.

[21] Yoshua Bengio, Deep learning of representations for unsupervised and transfer learning, Proceedings of ICML Workshop on Unsupervised and Transfer Learning, 2012, pp. 17–36 http://proceedings.mlr.press/v27/bengio12a.html.

[22] Srinath S. Kumar, Dulcy M. Abraham, Mohammad R. Jahanshahi, Tom Iseley, Justin Starr, Automated defect classification in sewer closed circuit television inspections using deep convolutional neural networks, Autom. Constr. 91 (2018) 273–283 https://doi.org/10.1016/j.autcon.2018.03.028.

[23] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in Neural Information Processing Systems, 2012, pp. 1097–1105 https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[24] Olga Russakovsky, Jia Deng, Su Hao, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252, https://doi.org/10.1007/s11263-015-0816-y.

[25] Ross Girshick, Jeff Donahue, Trevor Darrell, Jitendra Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587 https://ieeexplore.ieee.org/document/6909475.

[26] Yang Li, Wenming Zheng, Zhen Cui, Tong Zhang, Face recognition based on recurrent regression neural network, Neurocomputing 297 (2018) 50–58 https://doi.org/10.1016/j.neucom.2018.02.037.

[27] Morten Kolbæk, Yu Dong, Zheng-Hua Tan, Jesper Jensen, Morten Kolbaek, Yu Dong, Zheng-Hua Tan, Jesper Jensen, Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks, IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP) 25, No. 10, 2017, pp. 1901–1913 https://arxiv.org/abs/1703.06284.

[28] Erik Gawehn, Jan A. Hiss, Gisbert Schneider, Deep learning in drug discovery, Molecular Informatics 35 (1) (2016) 3–14 https://doi.org/10.1002/minf.201501008.

[29] L. Minh Dang, Syed Ibrahim Hassan, Im Suhyeon, Arun kumar Sangaiah, Irfan Mehmood, Seungmin Rho, Sanghyun Seo, Hyeonjoon Moon, UAV based wilt detection system via convolutional neural networks, Sustainable Computing: Informatics and Systems, 2018, https://doi.org/10.1016/j.suscom.2018.05.010.

[30] John Arevalo, Fabio A. González, Raúl Ramos-Pollán, Jose L. Oliveira, Miguel Angel Guevara Lopez, Convolutional neural networks for mammography mass lesion classification, 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2015, pp. 797–800 https://ieeexplore.ieee.org/document/7318482.

[31] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, Stefan Carlsson, CNN features off-the-shelf: An astounding baseline for recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2014, pp. 806–813 https://arxiv.org/abs/1403.6382.

[32] Yaniv Bar, Idit Diamant, Lior Wolf, Hayit Greenspan, Deep learning with non-medical training used for chest pathology identification, Medical Imaging 2015: Computer-Aided Diagnosis, vol. 9414, International Society for Optics and Photonics, 2015, p. 94140V https://www.researchgate.net/publication/300151231_Deep_learning_with_non-medical_training_used_for_chest_pathology_identification.

[33] Luis Perez, Jason Wang, The effectiveness of data augmentation in image classification using deep learning, arXiv preprint arXiv:1712.04621, 2017 https://arxiv.org/abs/1712.04621.

[34] Christian Wolf, Jean-Michel Jolion, Object count/area graphs for the evaluation of object detection and segmentation algorithms, IJDAR 8 (4) (2006) 280–296 https://doi.org/10.1007/s10032-006-0014-0.

[35] L. Minh Dang, Syed Ibrahim Hassan, Suhyeon Im, Irfan Mehmood, Hyeonjoon Moon, Utilizing text recognition for the defects extraction in sewers CCTV inspection videos, Comput. Ind. 99 (2018) 96–109 https://doi.org/10.1016/j.compind.2018.03.020.

[36] Ray Smith, An overview of the tesseract OCR engine, Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), vol. 2, IEEE, 2007, pp. 629–633, , https://doi.org/10.1109/ICDAR.2007.4376991.

[37] Martin Thoma, Analysis and optimization of convolutional neural network architectures, arXiv preprint arXiv:1707.09725, 2017 https://arxiv.org/abs/1707.09725.

[38] Tan N. Nguyen, Chien H. Thai, H. Nguyen-Xuan, Jaehong Lee, NURBS-based analyses of functionally graded carbon nanotube-reinforced composite shells, Compos. Struct. 203 (2018) 349–360 https://doi.org/10.1016/j.compstruct.2018.06.017.

[39] Tan N. Nguyen, Chien H. Thai, H. Nguyen-Xuan, Jaehong Lee, Geometrically nonlinear analysis of functionally graded material plates using an improved moving kriging meshfree method based on a refined plate theory, Compos. Struct. 193 (2018) 268–280 https://doi.org/10.1016/j.compstruct.2018.03.036.