Computers in Industry 99 (2018) 96–109

Contents lists available at ScienceDirect

Computers in Industry

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

Utilizing text recognition for the defects extraction in sewers CCTV inspection videos

L. Minh Dang, Syed Ibrahim Hassan, Suhyeon Im, Irfan Mehmood, Hyeonjoon Moon*

Department of Computer Science and Engineering, Sejong University, Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 1 August 2017
Received in revised form 25 November 2017
Accepted 16 March 2018
Available online 30 March 2018

Keywords:
Multi-frame integration
Sewer
Text recognition
MSERs
CCTV video

ABSTRACT

This paper proposed a novel automated framework for analyzing and tracking sewer inspection close-circuit television (CCTV) videos. The proposed model mainly supports the off-site examination and quality management process of the videos and enables efficient revaluation of CCTV videos to extract sewer condition data. The study discusses an important module for any automated analysis and defect detection in CCTV video. It includes two main modules: text recognition and cracks extraction. In the first module, multi-frame integration (MFI) was applied to reduce the background complexity, time and computational requirements needed for the video processing. Then maximally stable extremal regions (MSER) was used on the grayscale channel and HSV channel to effectively detect all the text edges. Saturation color channel was also applied to verify the detected text line and remove false alarms. In the second module, by utilizing the text information on each frame, the operator's operation during the inspection is simulated which would indicate valuable clues about the location and severity of the cracks. The proposed methodology was validated using a set of video provided by the Korea Institute of Construction Technology.

1. Introduction

Our generation witnesses a data boom in digital contents such as text, sound, video or images. Besides, there is also a potential of video-analysis within the decoding of social and cultural patterns [20]. As a result, there is an emerging need for effectively turning these data into knowledge which can be utilized in automated decision support system. Data mining including several common tasks such as association rule learning [16], classification [18] and clustering [17] and machine learning [19] have been received considerable attention from research and development communities.

As part of an important social urban infrastructure, sewer pipelines received enormous investments from the governments. They function as water quality protection, public health protection, sanitation, rain and drain excess water from rain, melting snow which will then be diverted to rivers, lakes or seas. However, these assets are under serious threat because when the sewer pipe get older, more defects and cracks emerge on its surface [1]. These defects eventually lead to the sewage system malfunction. An appropriate maintenance for the sewer pipe networks is imperative to ensure their operating efficiency because malfunction could lead to serious complications to the sewerage treatment plan.

Closed-circuit television (CCTV) is the most prevalent technology for inspecting or locating underground structures and defects maintenance as the setup costs are low compared to other technologies such as sewer scanner evaluation technology (SSET), ground piercing radar (GPR) and infrared photo. The robot records a video of underground pipe's conditions as it travels through sewer mains between utility holes. However, it always requires an operator to control remotely, so it highly depends on their assessments and evaluation. Also, the inaccurate evaluation of the sewer condition is expected to happen because of the long-lasting inspection period could lead to operator's tiredness. As a result, off-site reviews using the recorded video is usually performed. Various computer vision and artificial intelligence approaches have been utilized to support evaluators in analyzing sewer pipe's malfunction on the extracted images from the video [2–4]. Moreover, ubiquitous computing has drawn public attention in recent years; it assumed common devices and objects are to be smart computing devices connected to the Internet and capable of communication with users, and the other similar devices [21,22]. The concept could also be considered for the sewer crack detection application, various type of sensors can be distributed throughout the sewer

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

https://doi.org/10.1016/j.compind.2018.03.020
0166-3615/© 2018 Elsevier B.V. All rights reserved.

pipe networks, each sensor can communicate with others and automatically stream the data to the server for analyzing and warn the user if any serious crack appears.

A text extraction system typically includes three steps: The first step is localization which focuses on locating text lines followed by enhancing the quality which aims to increase the text and background contrast and the text quality for better recognition. Finally, in the recognition step, optical character recognition (OCR) engines were used for recognizing the text. Although OCR engines work best for scanned documents, they do not generate satisfactory results when the image has poor resolution, low contrast or too complex background. There have been various approaches in text detection, two widely used methods are sliding window classification and connected component analysis (CCA). The connected component-based methods applied by Koo et al. in [5], they considered text information as a set of distinct connected components based on color similarity or spatial layout. On the other hand, Zheng et al. in [6] applied sliding window classification approach, the classifier was fed with the positive label windows which contained text, and they were further categorized into text areas by applying morphological operations. Existing methods did solve specific text detection difficulty to some extent but since one method worked well on the specific type of dataset but became ineffective on other types because the key issues in text detection are the background complexity. For video subtitle detection, multi-frame integration should be considered because it increases the detection rate. For example, Guo et al. in [7] deployed multi-frame corner matching to lower the impact of the background on the text. The usual approaches for these methods were that they applied text detection on several consecutive frames then used MFI to verify the text areas. Opitz et al. in [8] used Maximally Stable Extremal Region (MSERs) as features to extract text or non-text components and proved that the accuracy was greatly improved compared to original features; it worked best on images with a very complex background. Then, Turki et al. in [9] implemented MSERs in HSV space color and proved that using HSV color channel outperformed the original RGB color channel in detecting the text pixels candidate. This paper investigates MSERs intensified by extracting edges and connection to refine the text components generated by the mask.

Previous research on defects detection in sewer pipe focused on using computer vision techniques such as [10,11]. However, the sewer pipes were observed in the completely dark environment, uneven lighting and noises do cause some flaws undetected. Recently, Halfawy et al. in [12] presented the optical flow techniques to estimate the camera motion parameters using Lucas-Kanade optical flow algorithm which correctly extracted the frames contained defects. However, this method tracked the optical flow vectors of two consecutive frames so if the video is too long; it will take long processing time and affect the overall performance.

Based on the above analysis and considerations, a novel approach to extract robust text features from sewer CCTV's video was proposed. The main contributions include three aspects.

1. Utilizing multi-frame integration on the extracted frames to improve the text edges and simplify the background.
2. Applying many computer vision based methods to improve the text detection including MSERs detection in HSV color spaces and grayscale, used saturation color channel as a reference point for the text line verification.
3. Detecting and extracting sewer defects location based on simulating the operator's activities using subtitle information.
4. Contributing the images used for training Tesseract OCR and the trained OCR model.

The remainders of the paper are shown as followed. Section 2 introduces the text detection and recognition for sewer framework. The experimental results are shown in Section 3 and Section 4 gives the conclusions of the paper and some approaches to improve the framework in the future.

2. Methodology

As depicted in Fig. 1, the proposed method consists of five main sections. 1) In the multi-frame integration step, 30 frames were extracted per second from the input video; then frame averaging was applied to enhance text edge and reduce background complexity. 2) Preprocessing the image to improve the detection and recognition rate. 3) Text detection composed of two steps: text localization to find the text lines. After that, the text lines were verified, any false alarms will be removed. 4) Text recognition consists of two steps, improved the extracted text quality and trained Tesseract OCR to recognize specific text type. 5) After collecting all the frames textual information, the model can simulate the observer's actions when the defects appear in the frames by applying the defects extraction module which contains defects detection and defects extraction.

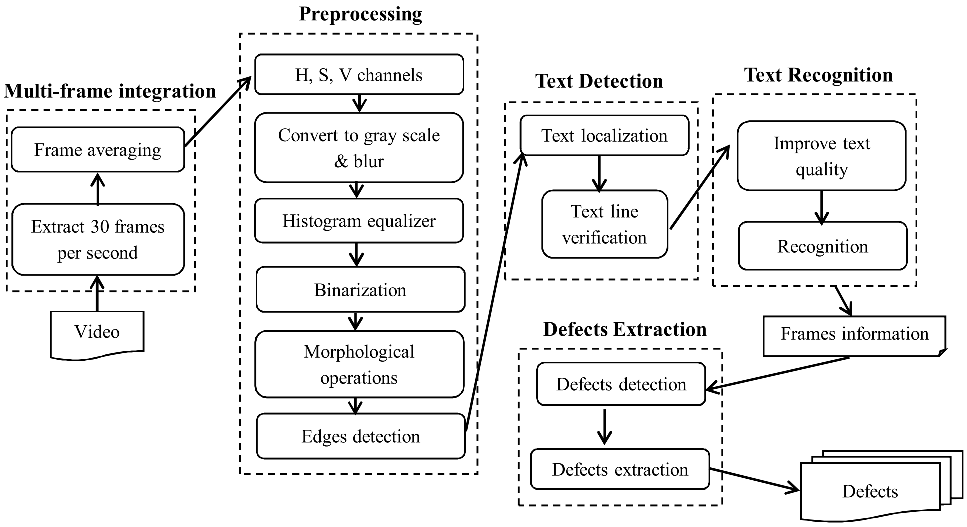


Fig. 1. The proposed architecture for sewer text detection and extraction.

2.1. Multi-frame integration

Although the CCTV video has many variants, the captions always appear in several consecutive frames where the background usually has a slight change as the robot moved in the underground pipe. The video is recorded at 30 Fps (frames per second). Thus, multi-frame integration (MFI) method was applied on a patch of 30 frames, and the MFI technique used in this study was multi-frame average as shown in Eq. (1)

$$Average\ Image_i(x,y) = \underset{j\in C_i}{avg}\left(p_j(x,y)\right) \tag{1}$$

let $p_j(x,y)$ is pixel value of frame j at position (x,y),C_i indicates frame cluster i contains frames between i and $i+29$.

Fig. 2 shows two examples of multi-frame averaging. Fig. 2(a) shows the frames before applying multi-frame integration technique, the background behind these images is quite complex and contains many edges which are easy to confuse the text detector. However, after applying the multi-frame averaging, the background complexity was vastly reduced as presented in Fig. 2(b).

2.2. Preprocessing

Because of the completely dark environment inside the drainage system, the robot was equipped with halogen lamps to brighten the environment. However, these lamps could reduce the recognition rate because of the uneven lighting condition in the video. Thus, preprocessing steps are mandatory to increase the detection rate of the text detector and recognizer.

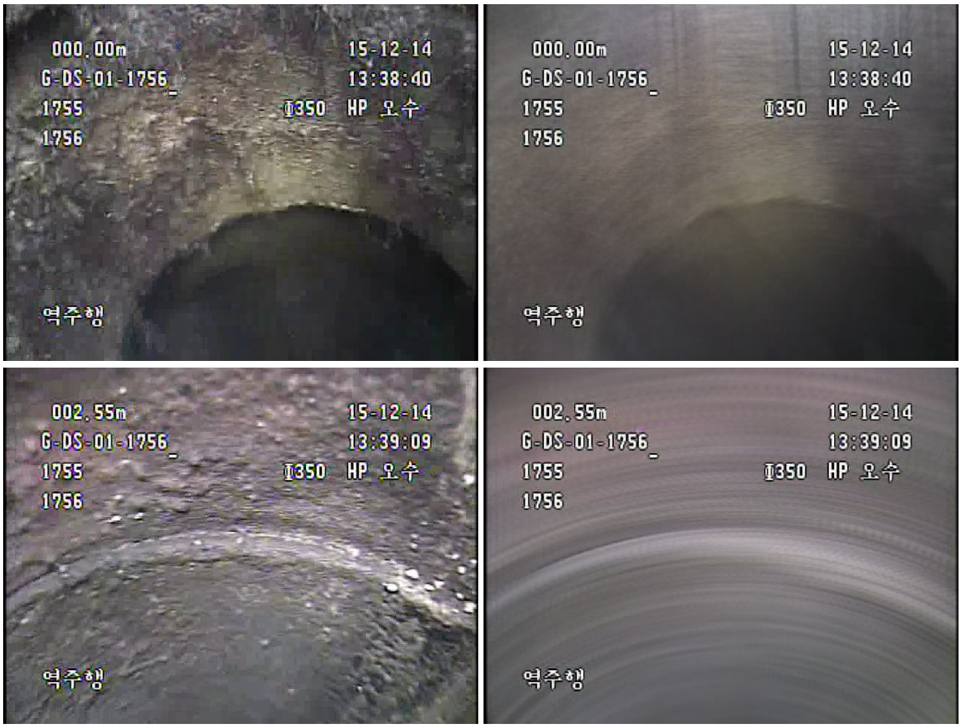
Firstly, the image was divided horizontally into two equal parts, and only the part which contained text was kept. After that, it was transformed into HSV (hue, saturation, value) color space to separate the color components from their intensity. Then each color channel was converted to grayscale, and Gaussian blurring using a 3×3 window was applied to remove noises of the grayscale image.

Since the images were captured under a dark environment, the histogram value for the low brightness was dominant. As a result, histogram smoothing was utilized to maximize the contrast by redistributing the brightness (brightness) distribution within the biased image and made it easy to extract the important features that were difficult to extract in pitch-dark images.

After that, they were converted into a binary image. The purpose of binarization is to categorize each pixel value into two classes. If the pixel value is higher than a minimum value, it is considered to be in one class (usually white); else it is given another class (usually black) because the images had varied illumination in different areas and contained noises in the caption area due to the lighting conditions. In this case, adaptive thresholding which calculated the minimum value for small image areas was applied. Separate thresholds were set for distinct areas of the image, and it obtained better outcomes for images with varying lighting.

Then morphological transformations were applied on these images, which include some operations based on the image's features. The structuring component contains a minor set of pixel neighborhoods which define formation and magnitude for the operator. Morphological operators examine the structuring components by fixing them into the items. For example, a structuring component which has particular extent and alignment was used to divide the text-like shape into binary photos. Formation, extent, and alignment properties of the structuring components can be chosen using the image object's properties. Primary morphological operators are merged to generate sophisticated operations like edge, contrast improvement, and image division. Two most commonly used operators are erosion and dilation. As defined in (2) and (3),

$$[\varepsilon_B(f)](x) = \min_{b\varepsilon B} f(x+b) \tag{2}$$



(a) (b)

Fig. 2. Sample images before and after applying multi-frame averaging: (a) Without multi-frame averaging. (b) With multi-frame averaging.

$$[\delta_B(f)](x) = \max_{b\varepsilon B} f(x+b) \quad (3)$$

with $\varepsilon_B(f)$ is the abrasion of image f by using structuring component B while $\delta_B(f)$ denotes the image f enlargement using structuring component B.

Finally, edge candidate detection was applied because the text lines had more edges than the background so horizontal and vertical edges for each channel were discovered by applying Sobel edge detection. After that, horizontal and vertical edges in each channel were merged to intensify the connection between edges which leads to the detection rate the improvement. The area of detection in this step will be improved in the next step by using the masks based MSERs.

2.3. Text detection

After completing previous preprocessing steps, maximally stable extremal region (MSERs) was used for text detection. MSERs was originally suggested by Matas et al. in [13] to spot similar elements from two images which have a different viewpoint, MSERs discovered steady connected areas through a scope of thresholds. Almost all the characters despite poor resolution could be detected using MSERs. Recently, Turki et al. in [9] detected the multi-channel MSERs as character candidates. As depicted in Fig. 3, this research explored the advantages of applying MSERs in HSV space color, and grayscale to refine the coverage text area and any false alarms will be removed.

2.3.1. Text localization

- MSERs masks in HSV color space

The result after completing previous preprocessing steps was three binary edge images for three color channels. They became the input for the following two processes. In the first process, three binary edge images were merged to create a combined binary edge image. During the second process, MSERs was used to detect text blobs in each binary edge image because it is important to detect text edge in multi-channel of space color.

Finally, binary edge images and MSERs in HSV color channels step were merged to create a combined edges image. However, some letters were not detected as some regions were unstable, so MSERs on the grayscale level was used to achieve a higher possibility to detect more characters' candidates.

- MSERs mask in grayscale level

As pointed out by [9], MSERs was very good for grayscale text detection so to detect more text regions that previous MSERs on HSV channels missed, MSERs has used again on the gray level image.

At the end of this step, the pixels between the combined edges image and MSERs in the grayscale level image were intersected to get only the text region pixels. Then bounding boxes around each region is reflected back to the original image as the final localization result in Fig. 4(c).

2.3.2. Text line verification

After completing the text localization step, all possible text lines were found. However, in Fig. 4(c), many noises were detected mainly due to the illumination and lighting conditions where the images were captured. Possible noises were removed using four conditions which validate the region as text or non-text as depicted in Table 1.

The above processing step did not cover all cases in the text line verification. Fig. 5(c) shows an example image which contains noise, but some are quite similar to the text lines, and we cannot

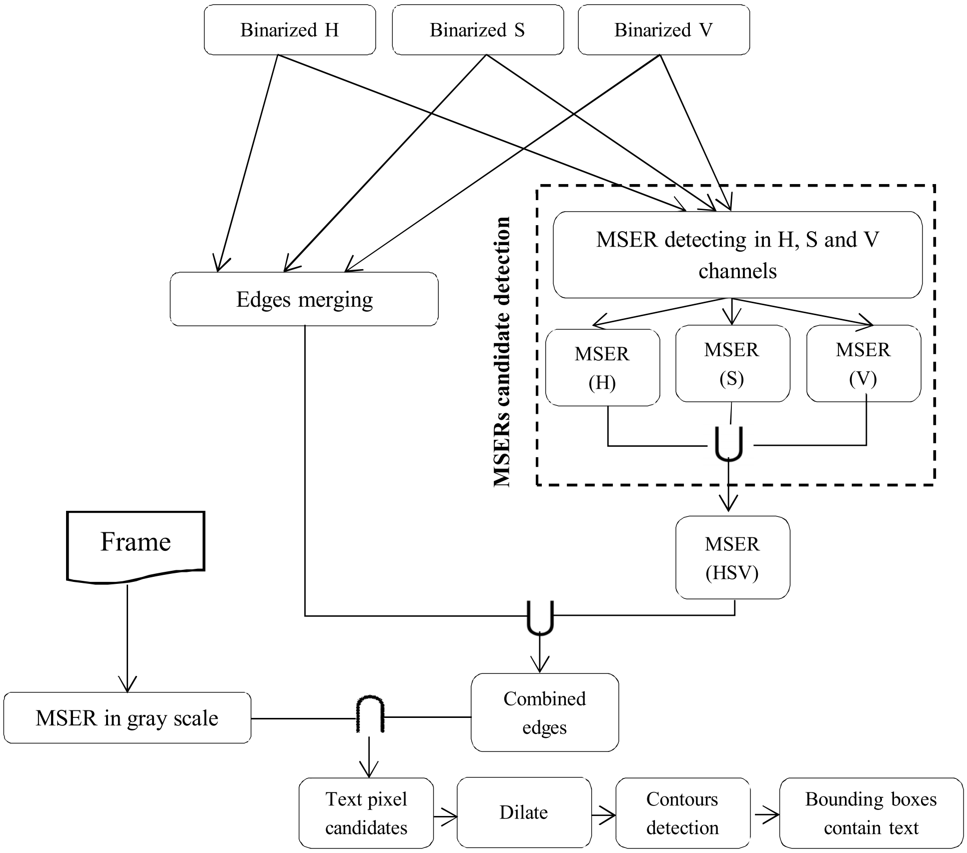


Fig. 3. Overall text localization process.

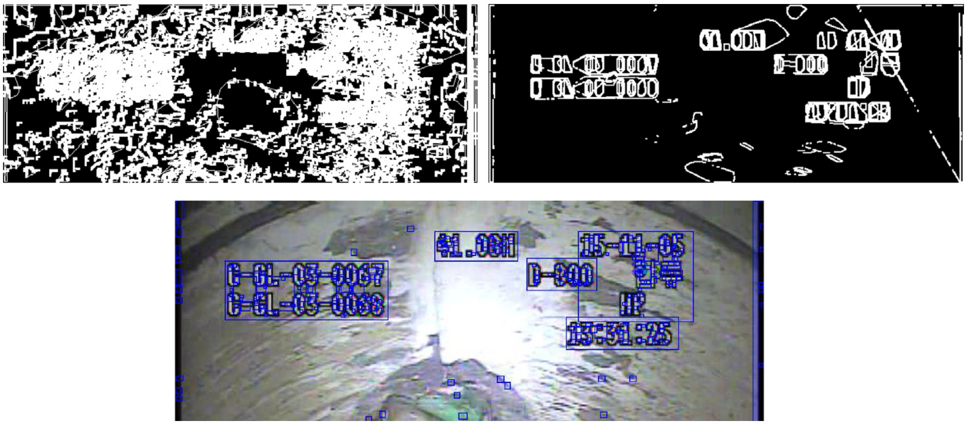
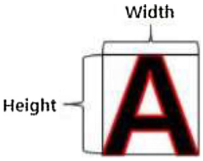
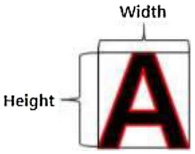
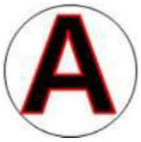


Fig. 4. Localization results. (a) MSERs in HSV merged with Edges detection in preprocessing step, (b) Result from intersecting the pixels of MSERs in grayscale and pixels of (a), (c) Final text localization result (blue bounding box). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 1
Text line verification conditions.

Name	Example	Condition
Contour size		Calculating the contour area of the text line. The valid contour area is between 10 and 5000.
Occupancy rate		Using the contour area of the text line, the width, and height of the surrounding rectangle. $Occupancy\ rate = \frac{Contour\ area}{Width*Height}$ The valid rate is between 0.0025 and 0.95.
Aspect ratio		Utilizing the width and height of the surrounding rectangle. $Aspect\ ratio = \frac{Height}{Width}$ The valid rate is between 1.5 and 25.
Compactness		Calculating the surrounding circle area and contour area $Aspect\ ratio = \frac{Contourarea}{Area\ of\ the\ circle}$ The valid rate is between 0.005 and 0.95.

remove them with just the above step. As a result, another processing step which used the saturation channel of the image was conducted, Fig. 5(b) shows the saturation channel extracted from the original image; it shows text region has more white pixels than the other region in the image. We used this to remove false text lines which had similar properties to the right one. The result before applying this method is shown in Fig. 5(c) and after applying the method is shown in Fig. 5(d).

2.4. Text recognition

In this research, the caption information from CCTV images is recognized by using the Optical Character Recognition (OCR) and template matching. Although the numbers of Korean letters in the subtitles were huge, the variation of them was limited; the Korean letters used in the video had a fixed number of characters so to increase the detection performance, template matching method was utilized. Since the numbers and type of alphanumeric characters were varying depend on the robot's configurations, the recognition phase was performed using the Tesseract OCR [14] engine.

2.4.1. Korean characters recognition

There were three types of Korean subtitles (project name, drainage type, location) in CCTV images as described in Table 2, each template is accompanied by its Canny edge detection result.

Template matching compares the template against the co-occurred spots size $w * h$. The similarity was calculated by the Normalized Correlation Coefficient matching method in Eq. (4) where I is the image, T refers to the template and R is the result. The sum-up is done for the image spot: $x' = 0 \ \ldots \ w - 1, y' = 0.h - 1$. The

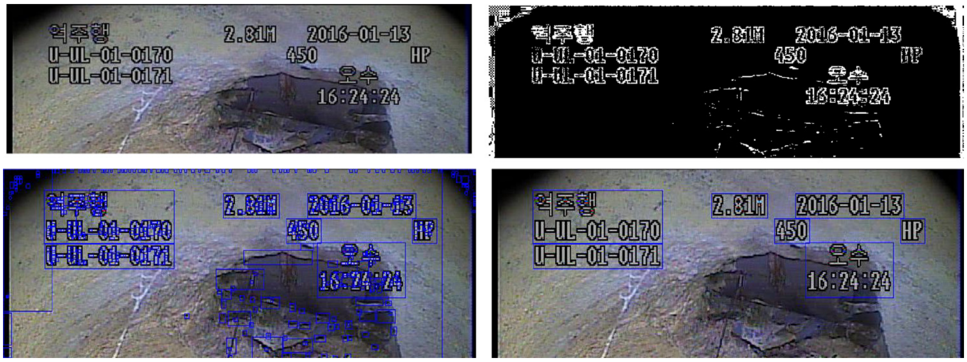
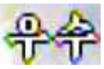


Fig. 5. Text line verification using saturation channel. (a) Original image, (b) The saturation channel (c) Text lines detection results (blue bounding box) and (d) Results after applying text lines verification. (For interpretation of the references to colour in this figure legend, the reader is referred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

Table 2
Example of Korean character templates.

English name	Template
Project name	노우하수관로정밀조사용역
Location	
Drainage type	
	

captions will be recognized when the matching degree of the extracted captions and the template is higher than the threshold value.

$$R(x,y) = \frac{\sum_{x',y'} (T(x',y').I(x+x',y+y'))}{\sqrt{\sum_{x',y'} T'(x',y')^2 . \sum_{x',y'} I(x+x',y+y')^2}} \tag{4}$$

There is one problem with the matching template is the sensitivity to the template and the image size, if the Korean character's size in the template were a little smaller or bigger than the Korean letters in the image, there would be a high chance that the template matching method could not match the letter. Thus, to overcome this problem, the multi-scale method was applied. It includes three steps:

- Iterating on the source image at numerous scales which aims to reduce the image size progressively.
- Utilizing template matching and following the match which has the highest correlation coefficient.
- Finally, select the area which has the highest correlation coefficient and labels it as "matched" area.

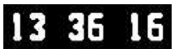
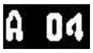
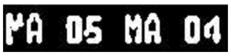
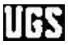
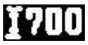
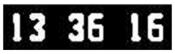
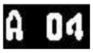
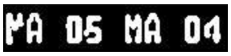
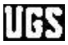
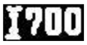
2.4.2. Alphabet and numerical recognition

Optical Character Recognition (OCR) is a leading-edge technology that enables text recognition by extracting text from a printed document or handwriting. In this research, the OCR was used for English and numerical caption recognition based on Google's Tesseract Application Programming Interface (API).

To recognize the subtitle information in the detected text lines, they had to be converted to a binary image. However, some images contain noises at the subtitle area due to the illumination, lighting conditions, etc. which lowered the recognition rate. In this case, fixed binarization is difficult if one threshold value was fixed for the entire image. Therefore, adaptive thresholding is applied to perform binarization on the corresponding text line. If the word is found in the page layout, it recognizes the word. On the other hand, when the recognition fails, recognition is repeatedly attempted through word pass 2 to improve the success rate. If the recognition is successful by performing the word pass 2, the word is recognized by fixing the height of the subtitle and the interval between the words and fixing Bigram between the words. Although Tesseract had its preprocessing step, it is extremely sensitive to background noise because adaptive thresholding is used for removing noise and background in the Tesseract API, and the subtitle recognition rate is remarkably degraded in case of a noisy image. Therefore, we applied preprocessing process to remove noise of sewer CCTV images by applying preprocessing steps as described in the previous section before feeding them into Tesseract. To improve the recognition rate of CCTV subtitling information of sewer pipes, OCR engine was trained by two types of font including 250 images for the first type and 400 images for the second type with the background removed.

Table 3 shows that before the font training, there were many cases in which the background and characters were well separated

Table 3
Example of recognition results after learning font information.

	Image	Recognition Result
Before learning font information	 13 36 16	13 86 16
	 A 04	4 04
	 MA 05 MA 04	14 06 M4 04
	 UGS	1165
	 I700	1 700
After learning font information	 13 36 16	13 36 16
	 A 04	A 04
	 MA 05 MA 04	MA 05 MA 04
	 UGS	UGS
	 I700	I 700

but mistakenly recognized the character to other characters. However, after training the Tesseract with two font types, almost all the characters were recognized correctly.

The Tesseract training process includes the following steps.

- Creating the box file

The file contains all the letters in the training file, in sequence, each letter per line along with the bounding rectangle's coordinates of the letter.

- Generating a set of all possible characters

Based on the character information processed by the boxing process, a set of all possible characters is generated then a learning dataset is constructed.

- Prototyping through clustering

From the constructed learning image, a prototype of the outline feature prototype of all characters and the number of features expected by the characters are generated.

- Generating directed acyclic word graphs

Finally, a dictionary of frequently appearing words is registered, and the subtitle font learning of the CCTV image is completed.

2.5. Defects extraction

After the text detection and recognition step, the textual information for each frame is extracted. By utilizing it, the observer's reaction when the defect appears can be simulated and tracked.

2.5.1. Defects detection

When the operator inspects potential cracks inside sewer pipeline, if the defect is detected, the robot stops moving or moves a little backward, then the mounted camera attached on the robot was used to inspect the cracks at various angles carefully. As a result, the "Travel distance" information remains the same (red box) at a period (usually over 3 s) as shown in Fig. 6.

However, due to possible noises, the recognition module can recognize wrong "Travel distance" or at worst completely misrecognize it. So it is very difficult to depend solely on the text recognition module for defects detection. As a result, this study proposed a new method for defects detection, firstly, all the frames in the video were extracted then on each frame the distance information was recognized. Also, the

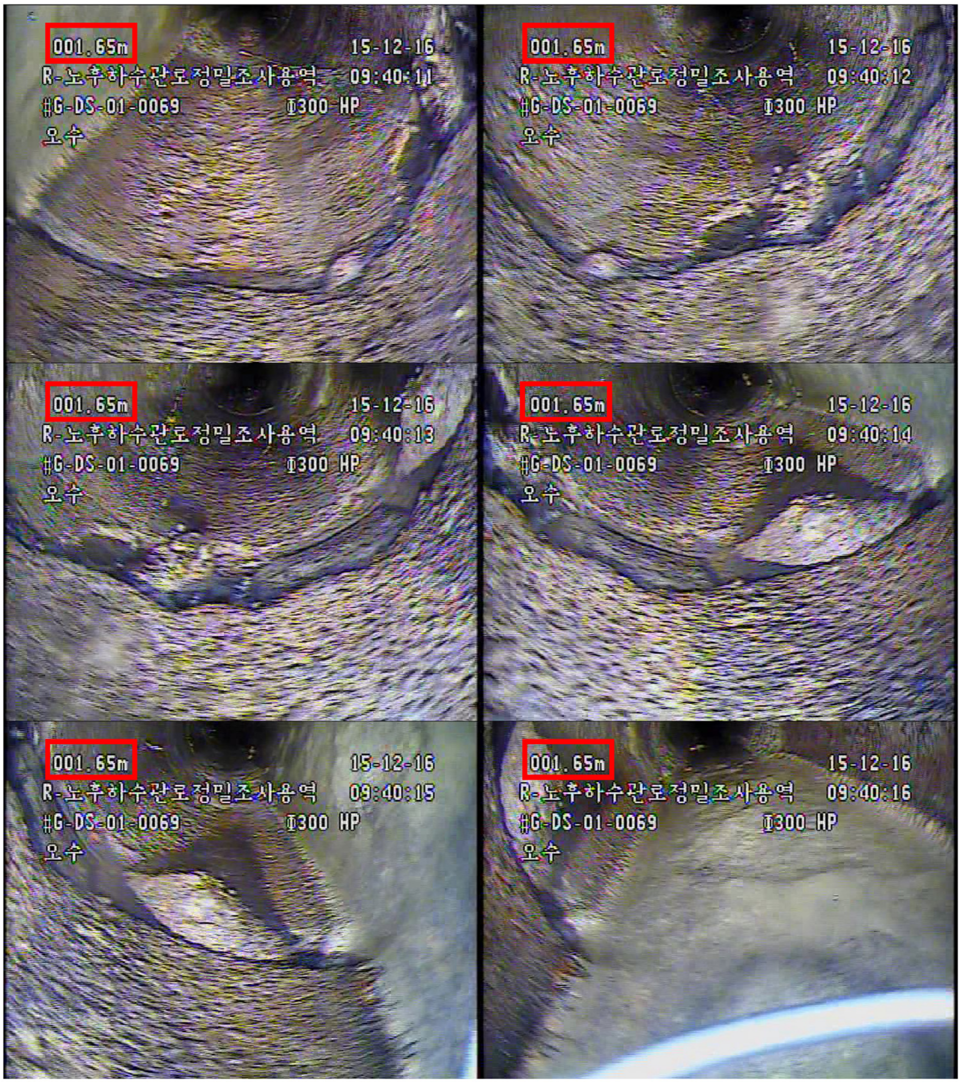


Fig. 6. Sample detected crack example; the robot stops for 6 s from (09:40:11 to 09:40:16). The red box indicates the "Travel distance" is the same. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

distance begins, and end position was calculated by using OpenCV CAP_PROP_POS_MSEC variable which returns the frame location in milliseconds. Although the recognizer may misrecognize the text information in some frames, other frames still be recognized correctly, so that is the reason for calculating begin position and the end position for each distance if the robot stops for more than 4 s which means the period between beginning and stop position must greater than 4000 milliseconds. The information including distance, the distance begin position and end position were extracted as shown in Algorithm 1. The output of the algorithm is a list of crack for extraction in the next step.

Algorithm 1. *Defects detection*

```
Initialize the parameters crack_collection;
For distance x in distance_collection do
    start_position = getBeginPosition(x)
    stop_position = getStopPosition (x)
    If stop_position - begin_position > 4000 then
        Add x, begin_position, stop_position to crack_collection
    End If
End For
```

2.5.2. Defects extraction

In this step, each crack in the crack collection from previous step will be extracted by using algorithm 2. At the end of this step, a list of frames ***f*** contained defects are extracted.

Algorithm 2. *Defects extraction*

```
For crack c in crack_collection do
    Set video start time to start position of c
    While start time < stop position of c
        f = read frame from video
        Extract f
    End while
End for
```

3. Experimental results

3.1. Dataset and evaluation method

In this research, we used the sewer CCTV video obtained from the Korea Institute of Construction Technology. These records were taken by the robot, to inspect the underground sewer pipes line. The duration of the video is from 1 to 15 min, whereas the subtitles are printed on the CCTV video, they contain crucial information for further investigation as shown in Fig. 7 such as travel distance, sewer pipe ID, date and time of inspection, type of sewer.

3.1.1. Image acquisition

Underground sewer pipes images were acquired by utilizing a commercial Robo Cam 5 (Tap Electronics Ind. Co., Ltd). It is equipped with 1/3 in. SONY Exmor CMOS camera module which captures 360° endless rotations and 240° side view up/down tilting. It has powerful halogen lamps to capture the images/videos in varying lighting condition. This robot is used to identify and locate any damages in underground pipes.

3.1.2. Dataset details

Subtitles in the sewer CCTV video/image includes the type of sewer pipe, location of the image acquisition, and the diameter of sewer pipe. They are the most crucial information for the investigation and management of the sewer pipe condition.

The proposed method was evaluated by two custom dataset extracted from sewer CCTV video; the detail description is shown in Table 4.

The evaluation protocol used in this study was recommended by Wolf et al. in [15]. The approach shows the object level precision and recall using detection quality restraints. Both amount and feature of detected bounding box matches were calculated for entire testing data. The assessment is calculated by Precision, Recall, and F-measure as followed:

$$Precision = \frac{\sum_N^i \sum_{|D^i|}^j M_D\left(D_j^i, G^i\right)}{\sum_N^i |D^i|} \tag{5}$$

$$Recall = \frac{\sum_N^i \sum_{|G^i|}^j M_G\left(G_j^i, D^i\right)}{\sum_N^i |G^i|} \tag{6}$$

$$F_{measure} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

where N is the number of data. $|D^i|$ and $|G^i|$ are the amount of detected and ground truth rectangles in image i-th. $M_D\left(D_j^i, G^i\right)$ and $M_G\left(G_j^i, D^i\right)$ are the matching scores for detected rectangle D_j and ground truth rectangle G_j. Two rectangles are judged as equal when their intersection proportion is greater than a fixed threshold, which manages the matching quality. The threshold for one to many matching was set to 0.8 for simple background dataset and 0.6 for complex background dataset because it is hard to detect the bounding box correctly.

The recognition performance was measured by the number of correctly recognized letters; it is defined as:

$$WRA = \frac{|C|}{|T|} \tag{8}$$

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

Table 4
Dataset description for text detection and recognition.

Dataset	Training	Testing	Total
Simple background	250	100	350
Complex background	400	300	700

3.2. Text detection and recognition on simple and complex background dataset

3.2.1. Text detection

The proposed model was implemented on two types of the dataset to measure the performance of the text detection module. The dataset with simple background contained 100 images whereas dataset with a complex background and various lighting conditions included 300 images.

- Simple background dataset

As explained earlier, template matching was used to recognize Korean letters, so only alphabetical and number were to be processed. Fig. 8 shows some examples of the results after we applied the text detection module. The module successfully detected all the text lines in the image. However, the system hardly detected the letter “m” because its tail had no text border and the background behind was similar to the text itself so when binarization was applied, the model considered it to be the background.

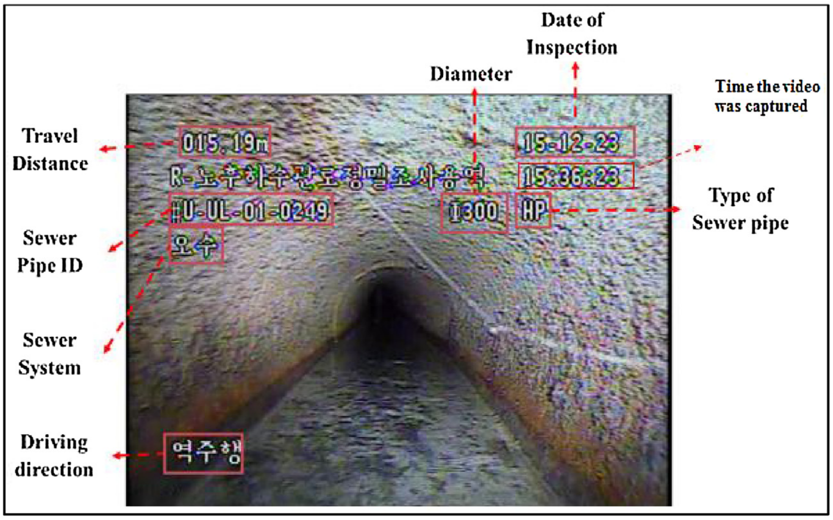


Fig. 7. Detailed description for the sewer CCTV video.

- Complex background dataset

Fig. 9 shows examples of the results on the complex background after applying the text detection module. It successfully detected the text lines in the image even though in some images the background was quite complicated.

Figs. 10 and 11 show the overall evaluation for the text detection module; we achieved high performance on both types of dataset. The dataset with a simple background and applied multi-frame integration method has 97% precision and 90% recall rate which lead to the high F-measure at 93% whereas the overall F-measure decreases to 90% without MFI. In case of the dataset with a complex background, the recall is 85%, the reason for the low recall is that many images in this dataset have the complex background or contains text which is similar to the background. Although various method had been applied, the model still misrecognized some text. However, the precision is at 87% which means the detector detect most of the text with a small number of false alarms.

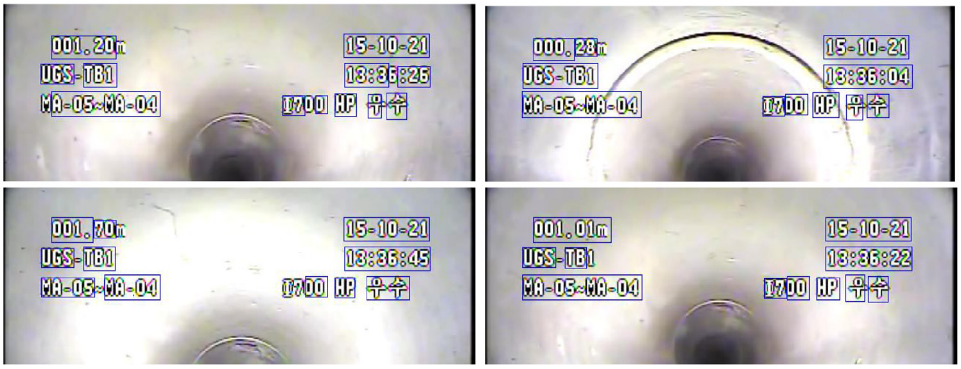


Fig. 8. Example of successfully detected text lines on the simple background (blue bounding box). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

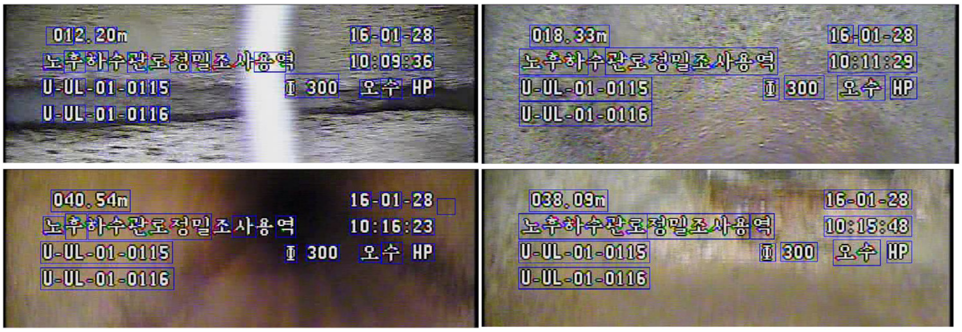


Fig. 9. Example of successfully detected text lines on the complex background (blue bounding box). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

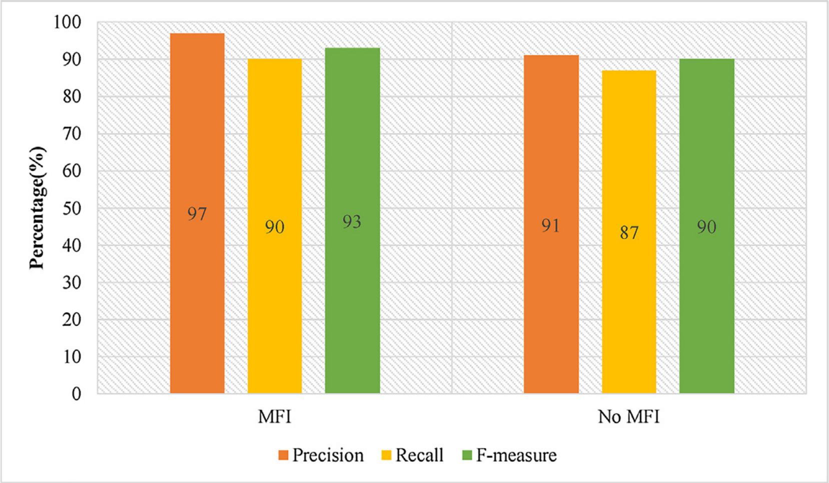


Fig. 10. Performance comparison between applying and not applying MFI on simple background dataset.

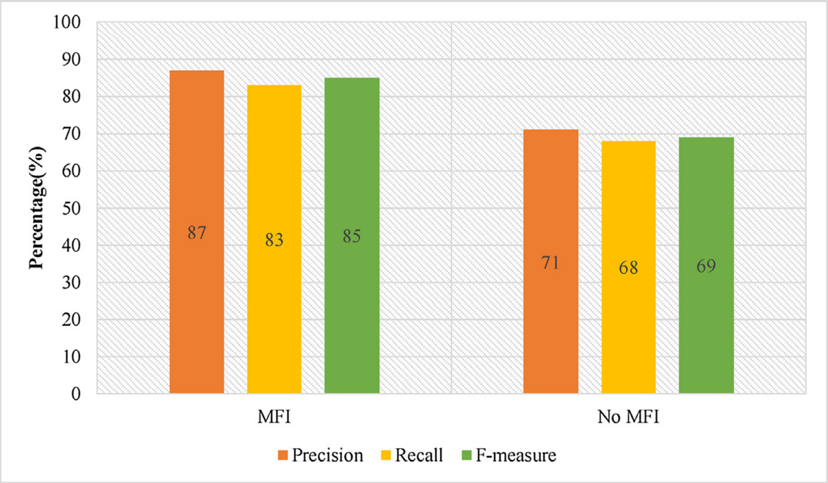
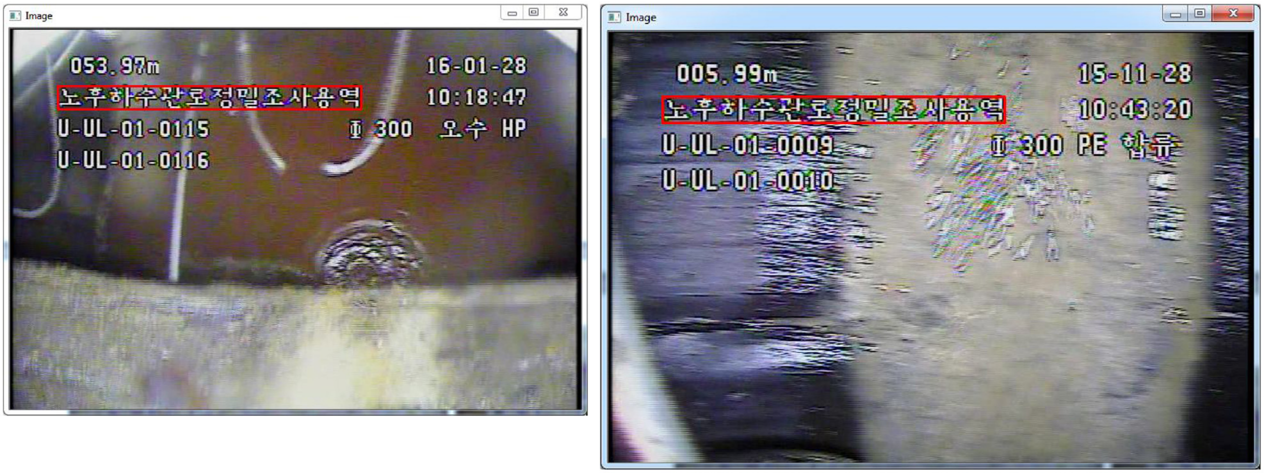


Fig. 11. Performance comparison between applying and not applying MFI on complex background dataset.



(a) (b)

Fig. 12. Template matching results on a dataset with complex background. (a) Image size (640 × 480) and (b) Image size (720 × 480)

Table 5
Accuracy for alphabet and number recognition on simple and complex background dataset.

Dataset	Accuracy
Simple background	94%
Complex background	87%

3.2.2. Text recognition on Korean letters

The threshold for matching on simple background dataset was 0.85 and 0.7 for complex background dataset. They were chosen based on the similarity and complexity between the background and text. Fig. 12 shows two examples of the template matching results on different image size (640 × 480) and (720 × 480). The matching method still worked well with the different image size and complex background. The matching result for Korean letter on simple background dataset was 100% while it was 93% on complex background dataset

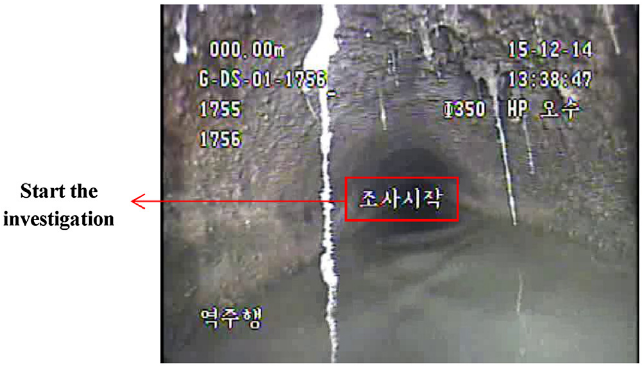


Fig. 13. Korean text for “start the inspection” in the video frame.

3.2.3. Text recognition on alphabet and number

Table 5 describes the recognition accuracy for simple background dataset and complex background dataset at 94% and 87%

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

Video ID	Video length (Seconds)	Total extracted frames	Frames after detecting “start of inspection.”	Total frames after applying MFI
1	682	20483	18234	607
2	675	20245	16681	556
3	418	12532	11125	370
4	303	9108	8312	277
5	526	15795	14963	498
6	680	20404	16325	544
7	563	16889	14823	494
8	597	17910	15121	504
9	639	19164	19164	638
10	1304	39099	39099	1303

Table 7
Sewer video subtitles detailed description.

Video ID	Inspection date	Sewer pipe ID	Diameter	Travel distance	Type
1	15-11-27	GDS012169 GDS012180	I400	000 m–> 052 m	HP
2	15-12-07	GDS011531	I700	000 m–> 047 m	HP
3	15-12-16	GDS010622	I300	000 m–> 027 m	HP
4	15-11-27	GDS010707 GDS010708	I300	000 m–> 041 m	HP
5	15-12-24	GDS012126 GDS012127	I300	0 m–> 41 m	HP
6	15-12-28	GDS012129 GDS012130	I600	0 m–> 63 m	HP
7	2015-11-28	GDS010283 GDS010284	I300	0 m–> 48 m	HP
8	2015-11-30	GDS010333 GDS010290	I300	0 m–> 42 m	HP
9	15-11-25	GDS011365	I400	0 m–> 47 m	HP
10	15-11-27	GDS011313	I300	0 m–> 58 m	HP

respectively. The recognition performance on the simple background was better than complex background dataset. Even though we used various preprocessing steps and trained the recognition model, some images had complex background and illuminations that the model could not recognize correctly.

3.3. Text detection on CCTV's recorded video

In addition to the experimental results above, detection and recognition system were further applied to ten videos to clarify the effectiveness of the multi-frame integration technique.

Because the robot started the recording on the ground before it was placed in the sewer underground, the model only needs to start extract frames when the robot was in the sewer. Thus, the frame which contains “start the inspection” subtitle as shown in Fig. 13 was searched. If it detected the “start the inspection” in one frame then all frames after the detected frame was extracted. On the other hand all video frames will be extracted.

All the videos were recorded at 30 frames/s on the different sewer system. The information regarding the length of the video, the total of 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.

The subtitle information is described in Table 7. In each video, only the travel distance change while other information stays the same for all the frames. We also select two videos (Video ID 7 and 8) as shown in Fig. 14 which have a different font than other videos to check if Tesseract OCR can recognize the text information.

The subtitle detection method was used on each video, and multi-frames integration was implemented. The ground truth labels were hand-generated, then they were used for comparing with the detected bounding boxes. The results were shown in Table 8.

Table 8 proves that using multi-frame average technique significantly reduced the wrongly detected boxes and simultaneously increased the accuracy. As the enhancement of the background quality improved the differences of low-resolution text, at the same time blurred the background which made text detection much easier. Also, the detection module can detect text boxes in the video which have small differences in text font and text format.

3.4. Defects detection and extraction on CCTV's recorded video

In this section, a comparison between the model defects extraction module and the report manually assessed and evaluated by the observer are conducted to test the model reliability. The same videos from the previous section were used for testing. Table 9 shows the number of defects / no defect by our prediction and in the report.

$$Accuracy = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} = 91\%$$

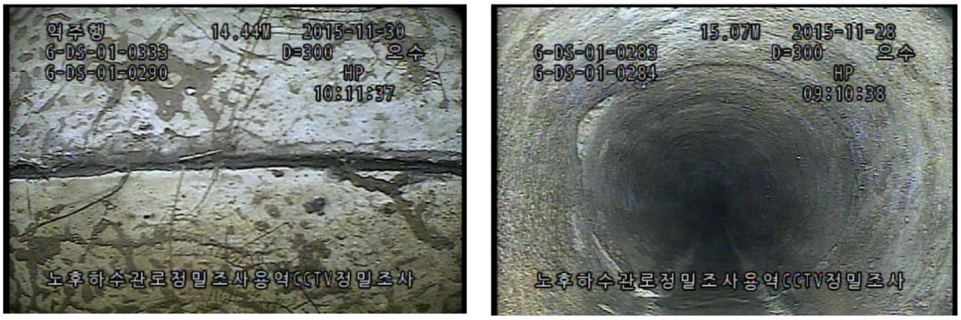


Fig. 14. Video which has different font and format than another video in the dataset.

Table 8
Text detection results on ten videos.

Total text boxes	Detected	False alarms
46328	41058	5270

Table 9
Sewer video subtitles detailed description.

		Prediction	
		Defects	No defect
Report	Defects	85 (TP)	6 (FN)
	No defect	33 (FP)	342 (TN)

$$Precision = \frac{T_P}{T_P + F_P} = 72\%$$

$$Recall = \frac{T_P}{T_P + F_N} = 93\%$$

The overall defects detection accuracy for ten videos is over 90%, which proves that the method successfully detects nearly all the defects comparing to the report, the recall rate is at 93% which indicates that only six defects are not recognized correctly. On the other hand, the precision is quite low at 72%, after carefully investigating the video and report, in some rare occasion, the robot stops, but there are no defects to check, it is due to the operator's control.

4. Conclusion

A novel text detection and recognition and defects extraction approach for underground sewer CCTV videos is introduced in this paper. Since most of the previous research have only focused on extracting the cracks by computer vision method, the proposed system focused on extracting them from textual information from sewer CCTV video. In text detection and recognition module, the combination of multi-frame integration, various processing steps and MSERs to extract the text edges make the method a truly robust one. The low false alarms rate will ensure the method provide accurate information for the real sewer analyst application. Moreover, this study also did the detection and recognition of Korean subtitles by applying multi-scale template matching method. The detection module obtained single-line text instead of text region contains multiple text lines, which benefit the recognition module. Although the study is designed mainly for detecting and recognizing text in sewer CCTV's videos, it works properly for most of the complex background videos. In the defects detection and extraction module, it utilizes the previous module and some unique attributes of the subtitle in the video to simulate operator's actions when the defect appears.

The proposed approach has been validated using a set of actual CCTV videos and images dataset extract from the sewer videos. For the simple and complex background dataset, the detection F-measure rate was at 93% and 77%, respectively while the recognition accuracy for simple background dataset was 94% and 87% for complex background dataset. On the other hand, for recorded sewer CCTV videos, the false alarm detected text lines after apply MFI was significantly reduced whereas the accuracy for defects extraction was at 91% for ten testing videos.

At its current implement status, the system serves two main purposes: (1) off-site checking and quality management process of the videos; and (2) enable efficient reevaluation of extracted videos to extract useful data. In the future, the system will be able to cope with live video streaming instead of recordings, thus assists the operators during the inspection process and overcomes issues related to operator fatigue and inadequate training.

Acknowledgements

This work was supported by Korea Institute of Planning and Evaluation for Technology in Food, Agriculture, Forestry and Fisheries (IPET) through Agri-Bio industry Technology Development Program, funded by Ministry of Agriculture, Food and Rural Affairs (MAFRA) (316033-4).

References

[1] Daeyoung Lee, et al., Study on evaluation technique for ground settlement by decrepit sewer using CCTV monitoring and GPR exploration in Korea, The 26th International Ocean and Polar Engineering Conference, International Society of Offshore and Polar Engineers, 2016.
[2] Mahmoud R. Halfawy, Jantira Hengmeechai, Efficient algorithm for crack detection in sewer images from closed-circuit television inspections, J. Infrastruct. Syst. 20 (2) (2013) 04013014.
[3] Iraky Khalifa, Amal Elsayed Aboutabl, Gamal S. Abdel Aziz, A new image model for predicting cracks in sewer pipes based on time, Int. J. Comput. Appl. 87 (9) (2014).
[4] Christian Koch, et al., A review on computer vision based defect detection and condition assessment of concrete and asphalt civil infrastructure, Adv. Eng. Inf. 29 (2) (2015) 196–210.
[5] 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.
[6] Yang Zheng, et al., A cascaded method for text detection in natural scene images, Neurocomputing 238 (2017) 307–315.
[7] Zhe Guo, et al., A method of effective text extraction for complex video scene, Math. Problems Eng. 2016 (2016).
[8] Michael Opitz, et al., End-to-end text recognition using local ternary patterns, MSER and deep convolutional nets, Document Analysis Systems (DAS), 2014 11th IAPR International Workshop On. IEEE (2014).
[9] 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. International Society for Optics and Photonics (2017).
[10] Mahmoud R. Halfawy, Jantira Hengmeechai, Automated defect detection in sewer closed circuit television images using histograms of oriented gradients and support vector machine, Autom. Constr. 38 (2014) 1–13.
[11] John Mashford, et al., Edge detection in pipe images using classification of haar wavelet transforms, Appl. Artif. Intell. 28 (7) (2014) 675–689.
[12] Mahmoud R. Halfawy, Jantira Hengmeechai, Optical flow techniques for estimation of camera motion parameters in sewer closed circuit television inspection videos, Autom. Constr. 38 (2014) 39–45.
[13] Jiri Matas, et al., Robust wide-baseline stereo from maximally stable extremal regions, Image Vision Comput. 22 (10) (2004) 761–767.
[14] Ray W. Smith, History of the Tesseract OCR Engine: What Worked and What Didn't, DRR, 2013.
[15] Christian Wolf, Jean-Michel Jolion, Object count/area graphs for the evaluation of object detection and segmentation algorithms, Int. J. Doc. Anal. Recognit. (IJDAR) 8 (4) (2006) 280–296.
[16] Bay Vo, et al., A novel approach for mining maximal frequent patterns, Expert Syst. Appl. 73 (2017) 178–186.
[17] Tung Kieu, et al., Mining top-k co-occurrence items with sequential pattern, Expert Syst. Appl. 85 (2017) 123–133.
[18] Minh Dang, Duc Duong, Improvement methods for stock market prediction using financial news articles, Information and Computer Science (NICS), 2016 3rd National Foundation for Science and Technology Development Conference On. IEEE (2016).
[19] Fu-Chen Chen, et al., A texture-based video processing methodology using Bayesian data fusion for autonomous crack detection on metallic surfaces, Comput.-Aided Civ. Infrastruct. Eng. 32 (4) (2017) 271–287.
[20] Lev Manovich, The science of culture? Social computing, digital humanities and cultural analytics, CA: J. Cult. Anal. 1 (1) (2016).
[21] Abolghasem Sadeghi-Niaraki, Bahman Jamali, Reza Arasteh, Application of geospatial analysis and augmented reality visualization in indoor advertising, Int. J. Geogr. Geol. 4 (1) (2015) 11–23.
[22] Abolghasem Sadeghi-Niaraki, Bahman Jamali, Reza Arasteh, Application of geospatial analysis and augmented reality visualization in indoor advertising, Int. J. Geogr. Geol. 4 (1) (2015) 11–23.

L. Minh Dang, He received BS degree majoring Information Systems in 2016 from the University of Information Technology, VNU, HCMC, VietNam. He is currently pursuing M.S. degree in Computer Science from Sejong University, Seoul, South Korea. He joined Computer Vision Pattern Recognition Laboratory (CVPR Lab) at the beginning of 2017. His current research interests include computer vision, natural language processing and artificial intelligence.

Syed Ibrahim Hassan, He received his BS degree in computer science in 2015 from Quaid-E-Awam University of Engineering Science and Technology Sindh, Pakistan. He is currently pursuing his MS degree in computer science and engineering from Sejong University, Seoul, South Korea. He joined CVPR Lab in September 2016. His current research interests include computer vision, deep learning, and image processing.

Irfan Mehmood received his BS degree in Computer Science from National University of Computer and Emerging Sciences from Pakistan. He is currently pursuing his Ph.D. degree at Sejong University, Seoul, Korea. His research interests include video summarization, medical image processing, and computer vision.

Hyeonjoon Moon, He received the B.S. degree in Electronics and Computer Engineering from Korea University in 1990. He received the M.S. and the Ph.D. degrees from Electrical and Computer Engineering at State University of New York at Buffalo in 1992 and 1999, respectively. From January 1996 to October 1999, he was a senior research in Electro-Optics/Infrared Image Processing Branch at U.S. Army Research Laboratory (ARL) in Adelphi, MD. He developed a face recognition system evaluation methodology based on the Face Recognition Technology (FERET) program. From November 1999 to February 2003, he was a principal research scientist at Viisage Technology in Littleton, MA. His main interest is on research and development is on real-time facial recognition system for access control, surveillance, and big database applications. He has extensive background on still image and real-time video based computer vision and pattern recognition. Since March 2004, he has joined the Department of Computer Science and Engineering at Sejong University, where he is currently a professor and chairman. His current research interests include image processing, biometrics, artificial intelligence and machine learning.