Defect Detection and Classification in Sewer Pipeline Inspection Videos Using Deep Neural Networks

Saeed Moradi

A Thesis

In the Department

of

Building, Civil and Environmental Engineering

Presented in Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy (Building, Civil and Environmental Engineering) at

Concordia University

Montreal, Quebec, Canada

July 2020

CONCORDIA UNIVERSITY

SCHOOL OF GRADUATE STUDIES

This is to certify that the thesis prepared

By: **Saeed Moradi**

Entitled: **Defect Detection and Classification in Sewer Pipeline Inspection Videos Using Deep Neural Networks**

and submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY (Building Engineering)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____Chair
Dr. Youmin Zhang

_____External Examiner
Dr. Mohammad Najafi

_____External to Program
Dr. Hassan Rivaz

_____Examiner
Dr. Ashutosh Bagchi

_____Examiner
Dr. Rebecca Dziedzic

_____ Thesis Supervisor
Dr. Fuzhan Nasiri

_____ Thesis Supervisor
Dr. Tarek Zayed

Approved by _____

Dr. Michelle Nokken, Graduate Program Director

July 8th 2020 _____

Dr. Mourad Debbabi,

Gina Cody School of Engineering and Computer Science

ABSTRACT

Defect Detection and Classification in Sewer Pipeline Inspection Videos Using Deep Neural Networks

Saeed Moradi, Ph.D.
Concordia University, 2020

Sewer pipelines as a critical civil infrastructure become a concern for municipalities as they are getting near to the end of their service lives. Meanwhile, new environmental laws and regulations, city expansions, and budget constraints make it harder to maintain these networks. On the other hand, access and inspect sewer pipelines by human-entry based methods are problematic and risky. Current practice for sewer pipeline assessment uses various types of equipment to inspect the condition of pipelines. One of the most used technologies for sewer pipelines inspection is Closed Circuit Television (CCTV). However, application of CCTV method in extensive sewer networks involves certified operators to inspect hours of videos, which is time-consuming, labor-intensive, and error prone.

The main objective of this research is to develop a framework for automated defect detection and classification in sewer CCTV inspection videos using computer vision techniques and deep neural networks. This study presents innovative algorithms to deal with the complexity of feature extraction and pattern recognition in sewer inspection videos due to lighting conditions, illumination variations, and unknown patterns of various sewer defects. Therefore, this research includes two main sub-models to first identify and localize anomalies in sewer inspection videos, and in the next phase, detect and classify the defects among the recognized anomalous frames.

In the first phase, an innovative approach is proposed for identifying the frames with potential anomalies and localizing them in the pipe segment which is being inspected. The normal and anomalous frames are classified utilizing a one-class support vector machine (OC-SVM). The proposed approach employs 3D Scale Invariant Feature Transform (SIFT) to extract spatio-temporal features and capture scene dynamic statistics in sewer CCTV videos. The OC-SVM is trained by the frame-features which are considered normal, and the outliers to this model are considered abnormal frames. In the next step, the identified anomalous frames are located by recognizing the present text information in them using an end-to-end text recognition approach. The proposed localization approach is performed in two steps, first the text regions are detected using maximally stable extremal regions (MSER) algorithm, then the text characters are recognized using a convolutional neural network (CNN). The performance of the proposed model is tested using videos from real-world sewer inspection reports, where the accuracies of 95% and 86% were achieved for anomaly detection and frame localization, respectively. Identifying the anomalous frames and excluding the normal frames from further analysis could reduce the time and cost of detection. It also ensures the accuracy and quality of assessment by reducing the number of neglected anomalous frames caused by operator error.

In the second phase, a defect detection framework is proposed to provide defect detection and classification among the identified anomalous frames. First, a deep Convolutional Neural Network (CNN) which is pre-trained using transfer learning, is used as a feature extractor. In the next step, the remaining convolutional layers of the constructed model are trained by the provided dataset from various types of sewer defects to detect and classify defects in the anomalous frames. The proposed methodology was validated by referencing the ground truth data of a dataset including four defects, and the mAP of 81.3% was achieved. It is expected that the developed model can help sewer inspectors in much faster and more accurate pipeline inspection. The whole framework would decrease the condition assessment time and increase the accuracy of sewer assessment reports.

Acknowledgment

I would like to thank my supervisor Dr.Tarek Zayed, for his guidance throughout completing this research and providing me with the opportunity to pursue my Ph.D. at Concordia University. I am also thankful for Dr. Fuzhan Nasiri, whom his presence has been a gift, and without his precious support, it would not be possible to conduct this research.

I would like to thank the faculty of Department of Building, Civil, and Environmental Engineering and in particular Dr. Ashutosh Bagchi for their continuous assistance. I am also grateful to my colleagues, friends, and committees who provided scholarships during my studies.

I also would like to express my gratitude to my parents and my family for their continuous support and encouragements. My sincere thanks also go to old friend, the best colleague, and mentor Ms. Farzaneh Golkhoo for her presence, as a source of energy, courage, and hope.

Dedicated,

To My Father and Mother,

To My Brother, and Sisters,

For All Help and Support.

TABLE OF CONTENTS

List of Figures

List of Tables

List of Acronyms

2D	Two-Dimensional
3D	Three Dimensional
ANN	Artificial Neural Network
AP	Average Precision
ASCE	American Society of Civil Engineers
CATT	Advancement of Trenchless Technology
CC	Connected Components
CCTV	Closed Circuit Television
CNN	Convolutional Neural Networks
CPU	Central Processing Unit
CTC	Connectionist Temporal Classification
DAE	Deep Auto-Encoder
DBN	Deep Belief Networks
DoG	Difference of Gaussians
FC	Fully Connected
FELL	Focused Electrode Leak Location
FN	False Negative
FP	False Positive
GAN	Generative Adversarial Nets
GLCM	Gray-Level Co-Occurrence Matrix

GPR	Ground Penetrating Radars
GPU	Graphic Processing Unit
HOG	Histograms of Oriented Gradients
IoU	Intersection over Union
LoG	Laplacian of Gaussian
LSCCR	Large Sewer Condition Coding and Rating
mAP	Mean Average Precision
Max	Maximum
Min	Minimum
MSED	Morphological Segmentation Based on Edge Detection
MSED	Mean Square Error
MSER	Maximally Stable Extremal Regions
NDT	Non-Destructive Testing
NRC	National Research Council of Canada
OCR	Optical Character Recognition
OC-SVM	One Class Support Vector Machine
P	Probability
PACP	Pipeline Assessment and Certification Program
PSET	Pipe Scanner and Evaluation Technology
RBF	Radial Basis Function
RBM	Restricted Boltzmann Machines
R-CNN	Region Based Convolutional Neural Network

RCP	Reinforced Concrete Pipe
ReLU	Rectified Linear Unit
ResNet	Residual Networks
RGB	Red-Green-Blue
RNN	Recurrent Neural Network
ROI	Region of Interest
RPN	Region Proposal Network
SIFT	Scale Invariant Feature Transforms
SRM	Structural Risk Minimization
SSD	Single Shot Multibox Detector
SSET	Sewer Scanning and Evaluation Technology
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
VCP	Vitrified Clay Pipe
WRc	Water Research Center
YOLO	You Look Only Once

Chapter 1 : Introduction

1.1. Background

Sewer collection system starts from civil and industrial outlets to laterals includes non-linear such as treatment plants, pumping stations, and lagoon systems, and linear facilities like sewer pipelines. Laterals convey the sewer medium to main pipes and interceptors, which then transfer sewage to treatment plants to provide primary, secondary, or tertiary treatment of wastewater. In the sewer system, pipelines make up a significant role since they expand all over the area to connect and deliver the sewer among the system elements. These vital networks are aging and reaching their service lives. In the US, it was projected that an average daily flow of around 50 million gallons of raw sewage is delivered to 19,500 sewer systems which are between 30 to 100 years old (Tuccillo et al. 2010). In Canada there are 143000 kilometers of sewer pipes that are equal to cross Canada 15 times from widest endpoints (Statistics Canada 2018). Despite environmental concerns of sewer pipelines, they are frequently neglected since they are buried and have low visibility. The continuous runoffs of sewer sanitary in the US reported at least 23,000 to 75,000 overflows, which results in the release of 3 billion to 10 billion gallons of raw sewer (Tuccillo et al. 2010).

Based on the American society of civil engineers (ASCE) report card for America's infrastructure (ASCE 2017), wastewater condition is graded as D (poor) condition in the US. According to the Canadian Infrastructure Report Card 2019 (FCM 2019), 18% of sewer pipelines in Canada are reported as poor to very poor condition and 17.3% in fair condition. Moreover, because of the degradation of sewer networks throughout their service life, even if they are in very good condition today, will require increasingly more substantial investments as they age (FCM 2019). Therefore, to prevent severe and costly damages, sewer system condition needs to be monitored through an appropriate and comprehensive periodic assessment (Guo et al. 2009b; Mohamed et al. 2019).

In summary, sewer pipelines suffer from poor condition ratings, and they are prone to failure and imposing costly consequences. Therefore, governments drive massive amounts of funds into wastewater system rehabilitation and improvements. Thus, it is necessary to conduct proper asset management and planning for existing wastewater pipelines and conduct developments in the system.

1.2. Problem Statement and Research Motivation

Proper and regular assessment of the infrastructures has to be done to evaluate the condition of the asset and consequently, deciding to rehabilitate or replacing the assets to have a cost-efficient operating system. Also, the regular condition inspection plays an important role in prolonging the estimated service life of the infrastructure. Regular assessments provide asset managers with enough data to look into the current situation of the asset and predict its future condition and make preventive decisions to avoid severe and destructive damages.

Sewer pipeline networks are one the vital infrastructures in cities as besides their primary service, any malfunctioning in their operation may affect the environment directly. Before the 1960s, inspecting sewer pipelines was a challenging task (Reyna et al. 1994), and the size of pipe made it

difficult for workers to access inside of the pipe for inspection. Thereby, innovative methods and technologies improved sewer inspection and assessment. Mechanical improvements in inspection technologies parallel to sensors and software developments offer fast and high-quality data acquisition. However, to select a suitable sewer inspection method, several factors such as pipe type, diameter, material, and cost need to be considered.

Currently, visual inspection using CCTV is the most widespread practice in sewer pipelines inspection and assessment. Visual inspection requires hundreds of hour data processing by certified operators to detect the defects (i.e., crack, joint offset, roots, deposit, infiltration, etc.) and assess the severity of defect (i.e., length, number, consequences, etc.). Moreover, recognizing the defects and assess their severity is subject to the operator's judgment. Based on the research conducted by Dirksen et al. (2013), 25% of defects are neglected by the operator during the inspection. Regarding the mentioned challenges, the main problems with manual visual inspection in assessing extensive sewer systems are that it is error-prone, subjective, and time-consuming.

In recent years, with the availability of powerful computers and advances in optical sensing technologies, application of computer vision techniques to automate sewer condition assessment has been an active research field. However, in previous studies, identification and localization of critical part have been done almost manually. One contribution of this research is to sense automatically the regions containing anomalies and potential defects and also being able to detect and classify defects. Employing the automated condition assessment can ease the current manual inspection and condition evaluation practice, which is labor intensive and time-consuming. Moreover, it increases the accuracy of inspection by reducing the number of neglected defects caused by operator fatigue or strain.

1.3. Research Objectives

The main objective of this research is to develop an automated tool to classify and detect the defects in sewer pipeline inspection CCTV videos to meet the following requirements: (i) the sewer inspection algorithm should be able to deal with CCTV videos in which illumination and video quality vary; (ii) the proposed model should be run in an automatic manner to minimize operator interference and user inputs; and (iii) the system should detect, localize, and identify defects with high consistency and accuracy.

Therefore, the objective can be decomposed into the following sub-objectives:

- Identify and study defect types and different pipeline characteristics;
- Develop an automated approach for feature extraction and anomaly detection in CCTV videos;
- Localizing the identified defected frame in the sewer pipe segment;
- Develop an automated defect detection and classification framework; and
- Implement the proposed framework on real-world problems to evaluate its applicability and performance.

1.4. Methodology Overview

After the statement and analysis of research problem in section 1.2 and identification of research objectives (Section 1.3), a comprehensive literature review has been conducted. In the literature

review, first, the most current inspection technologies are introduced, and their advantages and disadvantages are compared. The second part of the literature review inquires the potential use of an automated and reliable defect detection capability for sewer pipeline inspection and condition assessment through proposed models in recent studies. The literature review also introduces multiple object detection techniques that can be used in model development.

In order to develop an automated sewer inspection framework, different computer vision methods have been investigated for identifying regions of interest (ROI) and automatic defect classification and the following research questions need to be addressed:

- Which technologies are applicable in sewer pipeline inspection and which ones are the common practice?
- What are the current practices in video processing and computer vision technologies that can be employed in sewer pipeline assessment?
- What are the employed computer vision methods in automating defect detection in sewer pipeline assessment, and what are the achievements and limitations of each?
- Which modifications are required to justify and improve the application of a computer vision method for sewer defect detection?
- How to establish a generic methodology for automated defect detection in sewer pipelines considering variances in quality and illumination in CCTV inspection videos?

Taking into consideration the characteristics of sewer visual data (i.e., numerous defects, illumination variations, camera pose changes, etc.), most of the existing methods in image reasoning and pattern recognition are not applicable for defect detection. Therefore, it is critical to patch up inspection practices with the application of various computer vision and image recognition techniques.

Collaborative external partners such as The City of Laval and The Public Works Authority 'Ashghal' from Qatar, have provided the data used in this research. Videos and reports have been studied thoroughly to figure out real-world specific needs and inspection procedures. Models have been developed regarding the mentioned characteristics and validated through various statistical methods and also comparing experimental testing results against real inspection data.

1.5. Thesis layout

This study report has been organized based on the discussion in the research methodology (Section 1.3). Each chapter is intended to cover one of the research objectives and offset the limitations in the literature. Chapter 2 is the extended version of a previously published paper titled "*Review on Computer Aided Sewer Pipline Defect Detection and Condition Assessment*" in *Infrastructures* (Moradi et al. 2019a). In this chapter, related research about sewer pipeline inspection and condition assessment are reviewed and criticized. The chapter proceeds with the presentation of findings from the literature, the identification of research gaps, and the investigation of suitable techniques for the problem in hand.

Chapter 3 presents the research methodology in detail. In the proposed framework, the first section presents the anomalous frames recognition and localization among sewer defect detection videos. This section is a slightly modified version of a previously published paper titled "*Automated Anomaly Detection and Localization in Sewer Inspection Videos Using Proportional Data*

Modeling and Deep Learning–Based Text Recognition" published in *Journal of Infrastructure Systems* (Moradi et al. 2020).

Following in of chapter 3, a defect detection framework is proposed for defect detection and classification among the identified frames in the previous step. This section is a modified and extended version of the formerly published paper titled "*Automated Sewer Pipeline Inspection Using Computer Vision Techniques*" in ASCE Pipelines 2018 (Moradi et al. 2018a).

Chapter 4 starts with the description of a real-world case study and relative data collection. Then the introduced algorithms in chapter 3 are put into effect in the case study. The results are presented and tested against the available datasets to evaluate the generalizability of the proposed models.

Finally, Chapter 6 highlights the contributions, limitations of current research, and suggestions for future work.

Chapter 2 : Literature Review

Underground civil infrastructures always have been a concern for municipal asset managers since these utilities are exposed to unwanted and unpredicted environmental destructive causes such as decay, pressure, etc. Moreover, inspecting underground infrastructures is a demanding task due to accessibility problems. Sewer pipelines, as one the most vital infrastructures in modern cities, need to be assessed consequently. However, data acquisition and analysis is exhaustive and time-consuming, subjective to the operator's judgment, and full of human error. Dirksen et al. (2013) categorized the subjectivity of sewer pipelines assessment into defect detection, defect classification, and inspection interpretation. The authors found that a human operator misses 25 % of defects during the inspection process (Dirksen, et al., 2013).

In recent years, advances in visual and sensor technologies provide high-speed and high-quality data from sewer pipelines. Meanwhile, improvements in computer image and video analysis techniques have made automating sewer inspection and defect detection a point of interest for researchers. Several studies have been conducted through the application of various computer vision and machine learning algorithms on defect detection automation in sewer pipelines. In the following sections, these studies are introduced and criticized based on the used techniques by the researchers to discuss the shortcomings of each method and research gap.

Also, not all of the available inspection tools are applicable in the inspection of sewer pipeline, so a comprehensive comparison of the common technologies in sewer assessment is performed to discuss the advantages and limitations of each technology. This comparison may provide sewer inspectors a valuable insight to be able to choose the most suitable tool regarding various aspects of inspection such as pipe material, budget, etc.

This chapter is a slightly modified version of "*Review on Computer Aided Sewer Pipline Defect Detection and Condition Assessment*" published in *Infrastructures* (Moradi et al. 2019a) and has been reproduced here as the copyright is retained by the authors.

2.1 Defects in sewer pipelines

Various defects may affect sewer pipelines performance during their service life and shorten the intended pipe life span. The Pipeline Assessment and Certification Program (PACP) (NASSCO 2001), categorizes the existing defects sewer pipelines into three main categories: construction defects, structural defects, and operational and maintenance defects (Figure 2-1). Considering the environment of sewer pipelines, they are prone to be involved with a wide range of defects. These pipelines are installed underground, and the pipe can be disturbed by the stress of surrounding soil, traffic, water, vegetation roots, etc.

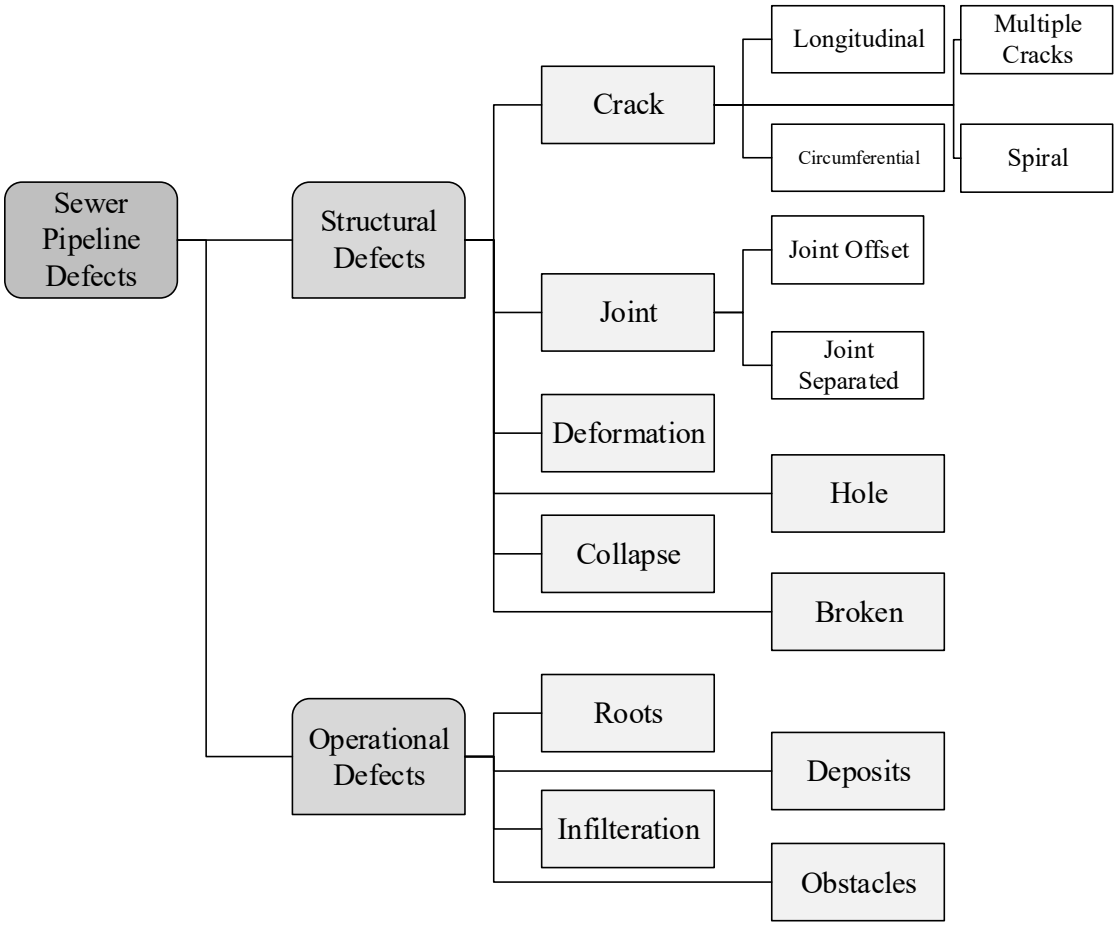


Figure 2-1. Sewer pipeline defect categories based on PACP

(Adapted from PACP (NASSCO, 2001))

Construction defects are generated during the pipe manufacturing and installation in excavated trenches. In this research, the defects emerging throughout service life of the pipe are studied, so only structural and operational defects are taken into consideration. In the next section different defect categories and a brief description for each of them are represented.

2.1.1 Structural Defects

The structural defects reduce the structural integrity of the pipeline and may result in structural failure. They mainly result from the external tensions on the pipe wall. Structural defects include cracks and fractures, deformation, collapse, breaks, and joint displacement. Different restoration decisions would be made based on the severity of the defects. Typically, in the early stages of the defect rehabilitation measures would be conducted. However, in more severe cases such as collapse or excessive deformation, the pipe needs to be replaced.

Cracks and Fracture

The cracks and fractures usually appear on the pipe walls. Cracks are lighter than fractures since the cracks in pipe walls are not distinctively broken apart while in fractures, the pipeline walls

become noticeably open. In both defect types, the pipe wall is still in place and does not fall apart. There are various types of fractures and cracks, including circumferential, longitudinal, multiple (complex), and diagonal. A longitudinal fracture or crack is parallel to the axis of the pipe. Circumferential is a fracture or crack that breaks in a circular plane perpendicular to the axis of the pipeline. A fracture or crack is considered spiral if it changes positions along the axis of the sewer pipe. A combination of the longitudinal, circumferential, and spiral defects in a relatively small area is considered as multiple cracks or fractures. Figure 2-2 shows different types of cracks in sewer pipes, and figure 2-3 shows a fracture in a sewer pipe.

(*a*) (*b*) (*c*)

Figure 2-2. Different types of the crack in sewer pipe: (*a*) Longitudinal crack, (*b*) Circumferential crack, (*c*) Spiral crack

Figure 2-3. Fracture in sewer pipe

Deformation

The deformation causes a reduction in cross-sectional area of pipeline that results in a decrease in flow capacity and surcharging of sewer sections. Deformation is measured as a percentage of the actual width (horizontal deformation) or height (vertical deformation) of the pipe that results in a noticeable change in the original cross-sectional area of the pipe.

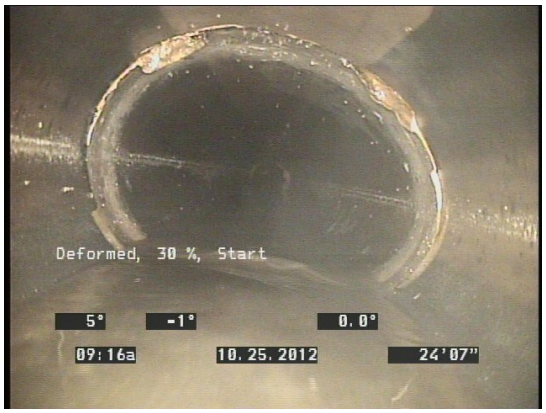


Figure 2-4. Deformation in sewer pipe (Adopted from (Iowa Great Lakes Sanitary District 2016))

Collapse

A collapsed pipe has lost its structural integrity, and half or more of the cross section is broken, and it is completely damaged and is out of service. A collapse defect has the highest level of criticality and requires immediate intervention since it stops the pipe operation in sewage transfer. Besides, pipe collapse results in exfiltration of sewer medium to the surrounding soil and contaminates underground water, which may cause serious health problems.

Figure 2-5. Collapse in sewer pipe

Breaks

The break is splitting and falling off the pipe wall material like small pieces, usually due to the expansion of corroded reinforcement or poor material, which generally is associated with fracture. Breaks are different to collapse since a broken pipe is localized, and the integrity of the pipe is not lost yet (Zhao et al. 2001). However, depending on the break location, it may cause infiltration or exfiltration.

Figure 2-6. Break in sewer pipe

Joint displacement

A displaced joint happens where the pipe is misaligned from its axis because of loading conditions, lack of lateral bedding supports, and construction problems. Displaced joint depending on water table may let infiltration or exfiltration and also increase the Manning coefficient that leads to rougher pipe internal surface and reduction in the hydraulic capacity of the pipe (Zhao et al. 2001).

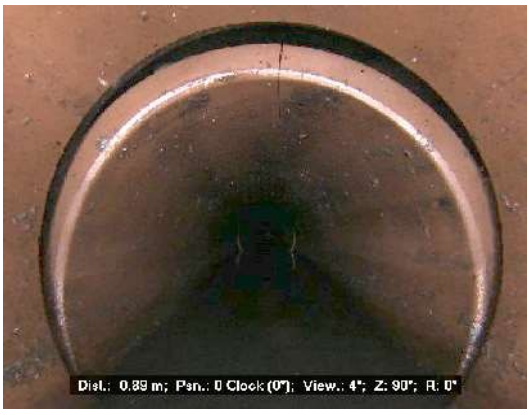


Figure 2-7. Joint displacement in sewer pipe

2.1.2 Operational Defects

Operational defects are all the defects that affect the operation and decrease the functionality of the pipe when conveying the flow. Operational defects, including infiltration, deposits, and root intrusion affect the serviceability of pipe. They usually are a result of structural defects such as cracks or joint displacement and can be cured by maintenance measures.

Infiltration

Infiltration is the incursion of groundwater into the sewer pipes due to displaced joints, holes, breaks, and physical damages. Sewer pipe infiltration can be graded as dripping, seeping, and

running. Moreover, exfiltration or leakage is the seeping of sewer flow out of the pipe through a specific defect. Both infiltration and exfiltration are damaging to the environment.

Figure 2-8. Infiltration in sewer pipe

Roots

Roots cause a reduction in the cross-sectional area of pipes and reduce the flow of the pipe. Pipes that have been laid until 5 meters deep from the surface and have plantation above them are more prone to root intrusion(Rahman and Vanier 2004). Roots penetrate from structural defects such as fractures, and holes leading to a reduced flow through blocking the pipes cross-sectional area. Also, when roots enter the pipe they start to grow and cause further structural defects.

Figure 2-9. Root intrusion in sewer pipe

Deposits

Another defect that may significantly disrupt the flow in sewer pipes is deposit. Attached deposits are the stuck materials on the pipe surface. While settled deposits are the remaining deposits on the pipe surface that could cause a reduction in pipe diameter. Deposit of silt is also called debris and, in some cases,, maybe a result of a piece of construction material (manufactural debris). It is a sign of more severe conditions in upstream (Rahman & Vanier, 2004).

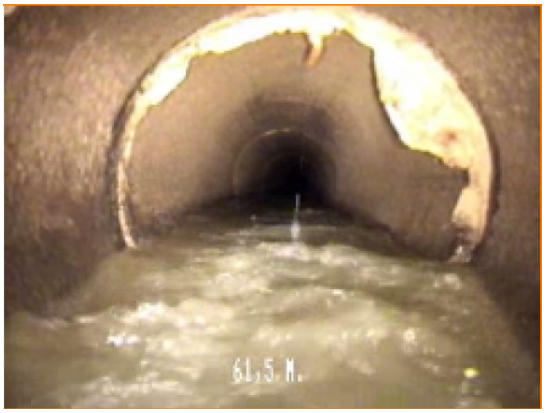


Figure 2-10. Deposit in sewer pipe

2.2 Sewer pipeline inspection technologies

To predict the degradation level and consequently decide on repair or replace/renew of a sewer pipe, it is required to assess it and inspect the existing defects (Najafi 2016). However, the hidden condition of underground infrastructures makes their inspection challenging. Human direct entry and inspection is unfeasible due to the extensive buried pipelines, small size of the pipes, and safety issues. Mentioned challenges were always key attributes to motivate the development of more complex inspection tools for sewer pipelines inspection. The improvements in sensor and lens technologies made it easier and faster to innovate and improve new detection techniques.

In this research, various technologies for sewer inspection are introduced and grouped into four categories. Visual technologies that are dependant on a CCTV camera to record the internal environment of sewer pipes. Structural and bedding inspection technologies that verify the pipe wall structural integrity and the condition of soil enveloping the pipe. Defect-specific technologies that can identify specific defect, and hybrid technologies which are combination of several tools (Figure 2-11). In the following, each category is illustrated entirely by describing sub-category methods, and the advantages and disadvantages of each are discussed. Finally, all the explained technologies are compared, considering different criteria.

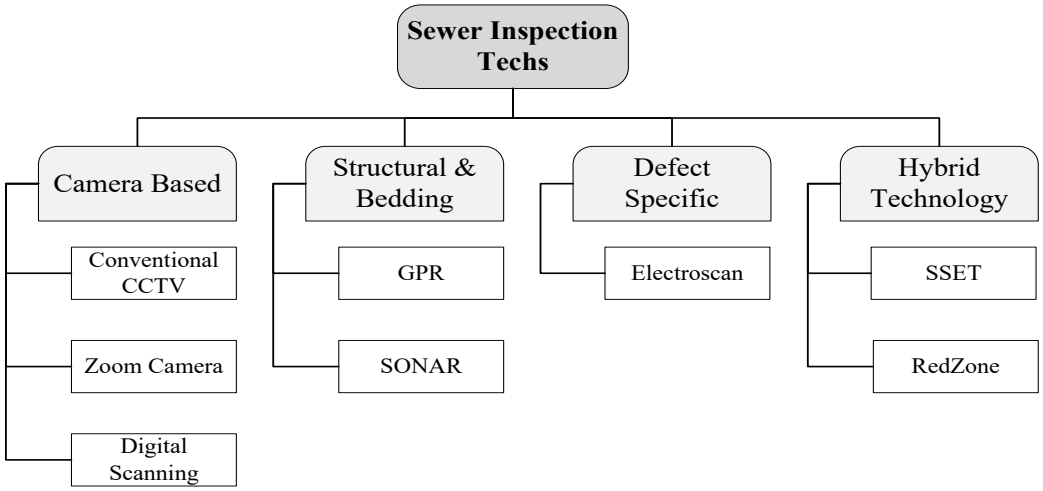


Figure 2-11. Sewer pipeline inspection tools (adapted from **(Moradi et al. 2019b)**)

2.2.1 Vision-based

In order to inspect the internal sewer wall, methods such as physical man entry or closed-circuit television cameras can be employed (Makar 1999). Man entry inspection is impractical and dangerous because of the sewer pipe's condition and environment. Therefore, camera-based inspection tools such as closed-circuit television (CCTV) inspection, zoom camera inspection, and digital scanning, are more applicable to assess the sewer pipelines visually.

The application of CCTV for the inspection of pipelines was first introduced in the 1960s. In this method, a camera is attached to a rover and an operator navigate it through the pipeline remotely. The distinct advantage of this method is that it provides evidence by directly illuminated images of pipe defects, which can be examined in detail by zooming the camera or viewing the defect from different angles by controlling the tractor (Hao et al. 2012). CCTV camera does not provide any data about pipe wall structural integrity or the soil condition surrounding the pipe and only provides the information about the pipe surface above the waterline (Selvakumar et al. 2014). On the other hand, CCTV technology is a productive and cost-effective tool that provides data of a wide range of sewer defect types.

Figure 2-12. CCTV inspection for sewer pipeline.

Image form IES (Undated)

Zoom camera is a camera attached to a retractable rod and performs manhole inspections. Like the conventional CCTV, the main application of zoom cameras is producing sewer pipe images and video footages. Zoom camera does not move through the pipe being inspected while it is fixed and placed through a manhole into the pipe. The main application of zoom cameras is to skim the pipelines. The pipe segment does not require to be cleaned, so the initial evaluation can be performed quickly to identify the segments for further inspection (Selvakumar et al. 2014). Therefore, zoom camera can be used to skim and prioritize the pipes for detailed inspection provided by CCTV camera.

The zoom camera inspection is productive, cost-efficient inspection method. However, there are some limitations as its application is limited to gravity sewers inspection since there is not any manholes access in force mains and service laterals. The same as traditional CCTV, zoom camera cannot inspect the pipe below the water surface. Also, if because of defects like sagging or deficient installation, the pipe deviates from a straight line, the hidden defects cannot be seen by the zoom camera. Moreover, the same detailed visual evaluation as conventional CCTV cannot be provided by zoom camera. Pan and tilt viewing is limited in some zoom cameras lack and the defects cannot accurately be measured or located. There are other limitations in image quality, illumination, and optical zoom.

Digital Scanning captures the pipe walls images using a 360-degree fisheye camera lens. A digital scanner examines each millimeter of pipe wall by high resolution images captured every ten centimeters and produces a constant image of the pipe. The recorded data is transferred to the surface station for real-time viewing and flagging for a more detailed evaluation (Tuccillo et al. 2010).

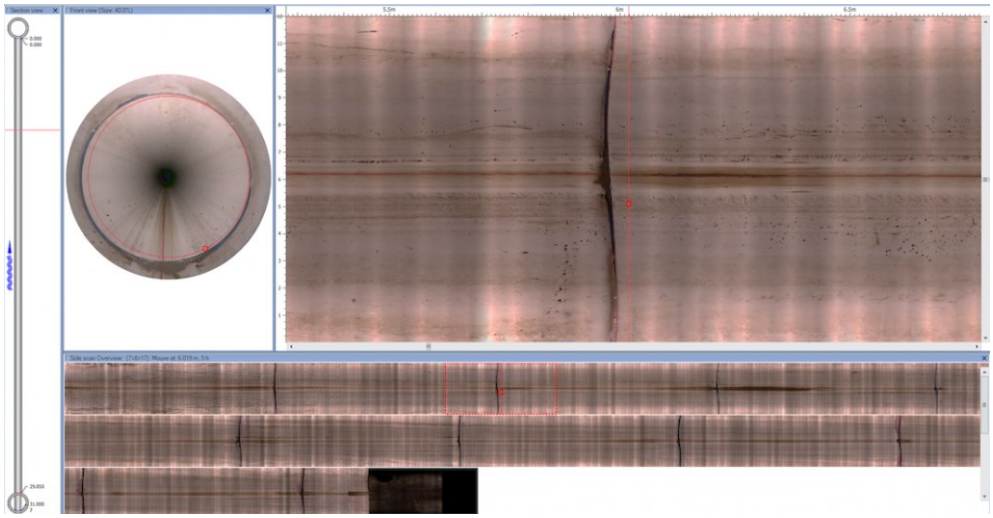


Figure 2-13. Digital side scanning

(Image adapted from Envirosight (Adams 2010))

Digital scanning can deliver high-quality images of the pipe wall in a shorter period. Data is also more appropriate for computer vision and image analysis applications. However, the main drawback comparing to conventional CCTV is cost efficiency. Digital scanning is relatively more expensive than CCTV.

2.2.2 Structural and bedding

Pipe wall integrity and bedding conditions cannot be inspected using visual technologies, so other technologies such as Ground-penetrating radar (GPR) have been used to examine subsurface conditions. Currently, GPR is the most applicable alternative to evaluate bedding and void conditions around the pipe wall. In addition, GPR as a non-destructive inspection method uses electromagnetic waves to evaluate subsurface materials (Hao et al. 2012). In GPR inspection, the location of pipes can be detected independent of pipe material. Therefore, precise data about the pipe wall and condition of the soil around pipe would be provided. However, magnetic pulses lose strength in conductive materials and ground material affects the penetration depth. Also, trained and certified operators are required to interpret the data provided by GPR (Tuccillo et al. 2010).

Another inspection tool for evaluating invisible areas is sonar. It functions by sending high-frequency sound waves through the pipe and signals vary based on the material condition of the pipeline. Sonar is able to detect defects located under the waterline as well as defects like joint displacement, and pipe deflection. Sonar does not need to shut down the sewer system. A sonar image is generated using the acoustic frequency. Considering parameters such as pipe diameter, amount of water sediment, and turbulence in the pipeline, various frequencies might be needed (Tuccillo et al. 2010). The provided reports are not straight forward need to be interpreted by trained operators.

2.2.3 Defect-specific

Defect specific technologies offer detection and severity of defects like infiltration and exfiltration. The electric resistance of the pipe wall defines the severity of infiltration or exfiltration. Electro

Scan Inc. introduced a tool to sense and measure infiltration defects in the sewer pipeline and based on the current flow the pipe wall integrity can be determined (O'Keefe 2013). The method can be used for pipes with non-conductive materials such as PVC, vitrified clay pipe (VCF), reinforced concrete pipe (RCP), and brick.

2.2.4 Hybrid

In recent years, to detect various types of defects in sewer pipelines, new inspection methods have emerged by combining different technologies. These methods tend to cover the limitations that are faced in other technologies. Sewer Scanner and Evaluation Technology (SSET) employs a fisheye camera lens combined with an optical scanner and gyroscope technology to present a total view of the pipeline surface. For further analysis, the provided images are digitized as color-coded computer images (ECT Team 2007). RedZone company co-operated laser and CCTV methods to inspect large diameter pipes. The tool provides more complete condition data along with information of the underwater defects like deposits (Guo et al. 2009a). Other multi-sensor inspection methods, such as KARO and PIRAT systems can detect and sort sewer defects automatically (Martel et al. 2011). INNOKANIS has introduced SewerBatt to cover up the limitations of CCTV tools. It incorporates zoom-camera and radio technology to benefit from optical and acoustic tools at the same time (Plihal et al. 2016).

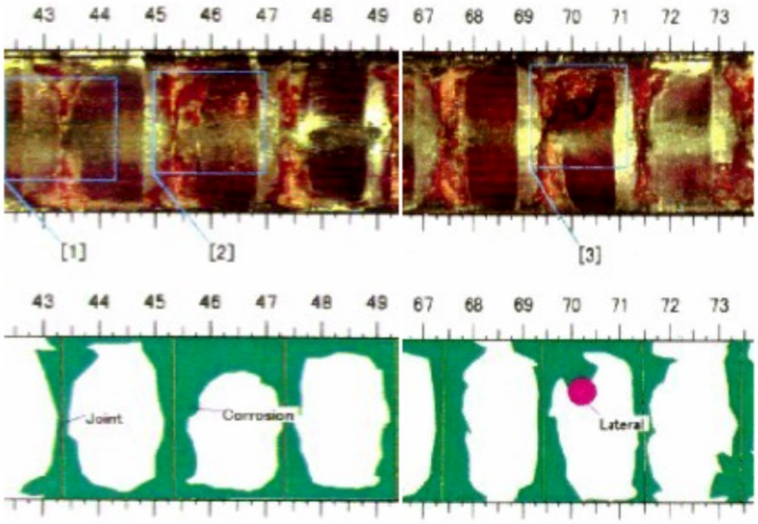


Figure 2-14. Sample image of sewer pipe –SSET

2.2.5 Sewer inspection technologies comparison

A broad range of inspection tools is now available for Municipalities. However, selecting the inspection technology relies on parameters like available budget, type of required assessment, and pipe material. Some studies presented a comparison between the various inspection tools regarding their benefits and drawbacks and proposed a broad outline of the existing practices and their future developments (Costello et al. 2007; Hao et al. 2012; Makar 1999; Selvakumar et al. 2014; Tuccillo et al. 2010; Wirahadikusumah et al. 1998). The advantages and limitations of introduced methods are described in the following.

As the most common tool in sewer inspection, camera-based technologies are able to provide visual evidence of most of sewer pipe defects. They are productive and cost-efficient, and the information is easy to analyze. However, they are limited in providing surrounding soil information and defects placed under the waterline. Also, the evaluation results are highly dependent on image quality, lighting, and illumination condition. Furthermore, the defects severity and characteristics such as cracks depth or the extent of deformation are subjective and rely on the judgment of the operator. Structural and bedding inspection tools can provide a cross-section profile of pipe wall and the condition of invisible parts such as the underwater line or outer pipe wall can be determined. The main limitations of these methods are the complexity of inspection data interpretation and required certain operational conditions.

Defect-specific tools are well established for the provided quantitative measures of the identified defect. In addition, the application is limited to the detection of one or two defects, costly to operate, and data interpretation requires trained operatives. Hybrid technologies the limitation of one tool is offset by employing two or more complementary tools, particularly camera-based methods. However, hybrid technologies are in the prototype phase and their operating expenses are still too high. Moreover, specialized preparation for running and data interpretation and supplementary equipment for fieldwork are required. Table 2-1 shows a comparison of the technologies mentioned above.

Table 2-1. Sewer inspection technologies comparison (adapted from (Moradi et al. 2019a))

	Vision Based		**Structural & Bedding**	**Hybrid Technology**	**Defect Specific**
	CCTV	Zoom camera	GPR	SSET	Electro scanning
No. of defects*	6	6	3	8	3
Complexity	Low	Low	Medium	High	Medium
Cost	Medium	Low	Medium	High	Medium
Downtime	High	Low	High	High	Low
Data analysis	Low	Low	Medium	High	Low
Operational equipment	Low	Low	Medium	High	Medium
Data quality	Medium	Low	High	High	High

* Defects that are inspected in sewer pipeline: deposits, debris, roots, sags and deflections, surface damage, joints displacement, cracks, infiltration, Bedding Condition.

2.3 Digital image

A digital image is an encoded representation of a real scene in an array of numbers as pixels in matrices which can be decoded by a computer. In an image, each matrix element is called pixel, which each pixel corresponds to an intensity value. Digital images can be interpreted in various forms, such as binary images, greyscale, and color images. Binary images present the image as 0 or 1. Greyscale images present the image in a range of 0 to 255 as the intensity level of gray in each pixel. A color image is a blend of several layers of intensities into one single matrix. The color images can be defined in numerous color spaces like RGB, CMY, HIS. For example, RGB color space blends three layers of red (R), green (G), and blue (B). So, in RGB color images each pixel matches up to three intensity values. Therefore, the techniques that are applicable to the monochrome images can also be used in processing color images (Gonzalez and Woods 2006).

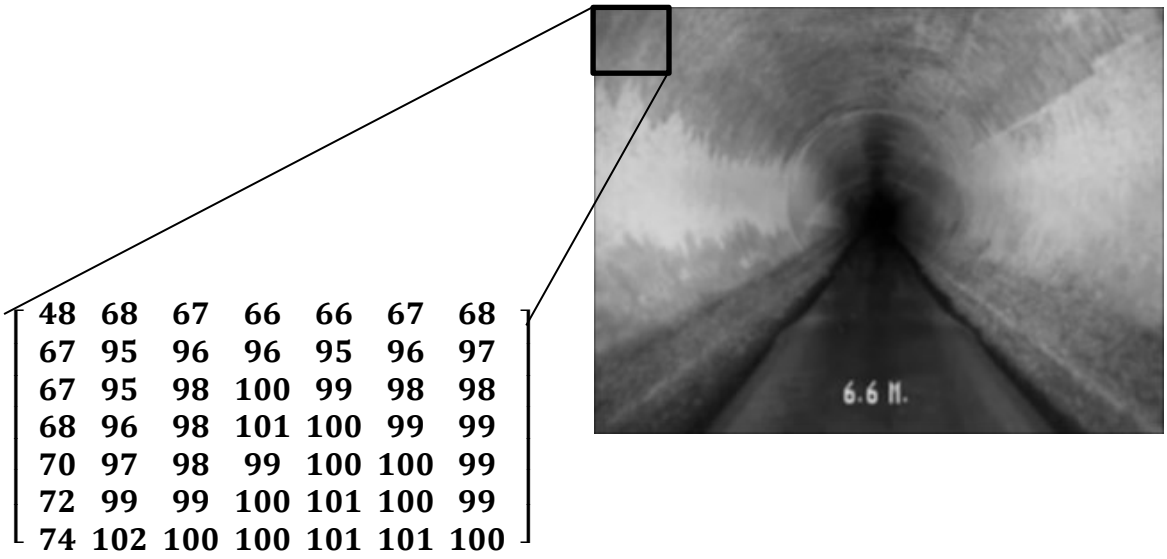


Figure 2-15. Representation of intensity values in monochrome image

This intensity matrix is the base for all image processing operations, and also image features are extracted due to values of the matrix using image processing techniques. Digital image processing is the analysis and interprets the characteristics of digital images through mathematical algorithms. Digital image processing is necessary to provide data for pattern recognition and object detection. All computer vision problems start with video recording and image acquisition. The next steps can be low-level image processing techniques, image segmentation, high-level algorithms, object detection, and data extraction. Low-level techniques include image preprocessing methods to enhance acquired images.

2.4 Image processing

Due to special conditions of the internal environment of sewer pipelines, the recorded videos and captured images are usually subject to artifacts and noise. Lighting conditions and illumination also affects the quality of images. Thus, some specific image processing operations as a preprocessing stage seems to be essential for input data preparation. Image preprocessing aims to remove distortions in an image to enhance the image quality. Also, some algorithms tend to enrich

image features like edges or apply geometric conversions like resizing and rotation on images (Sonka et al. 2007).

2.4.1 Image enhancement

Image enhancement algorithms alter the artifacts and noises in recorded videos because of imaging conditions and highlighting those image features which are making much of the characteristics of defects. Various filtering algorithms are applicable to digital images and can make them proper for computer image processing. Gaussian filtering is an effective 2D filtering algorithm to blur images and remove noise and undesired details in digital images. More image enhancement techniques can be found in (Jahne 2002).

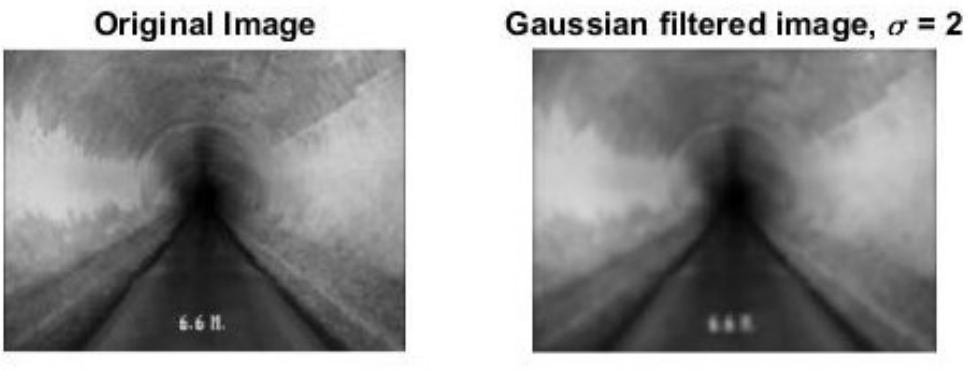


Figure 2-16. Example of Gaussian filtering

The Gaussian filter determines a weighted average of the neighborhood of each pixel in which the pixels that are farther from the central pixels assigned smaller weight so boundaries and edges can be defined clearly.

2.4.2 Morphological operation

Morphological operation is one of the main processing techniques performed on grayscale and binary images. These operations capture the structural elements of image objects based on their shape attributes (Qidwai and Chen 2009). Morphological operators determine the output image pixel values based on two elements: the condition specified by the set operator and by comparing the corresponding pixel in the input image with its neighborhoods (3×3 pixels). The process will then be applied over the whole image, and the pixels are compared with the array of underlying pixels. If two arrays of elements pixels are consistent with the set operator condition, then the central pixel of the neighborhood origin will be adjusted to a predefined value (Qidwai and Chen 2009).

The main morphological operators are dilation and erosion, closing and opening, and thickening and thinning. Each pair of mentioned operators performs opposite functions, respectively. Extending operators dilation, closing, and thickening make white elements in a binary image more predominant through various degrees. On the other hand, erosion, opening, and thinning reduce the size of the mentioned white elements (Qidwai and Chen 2009).

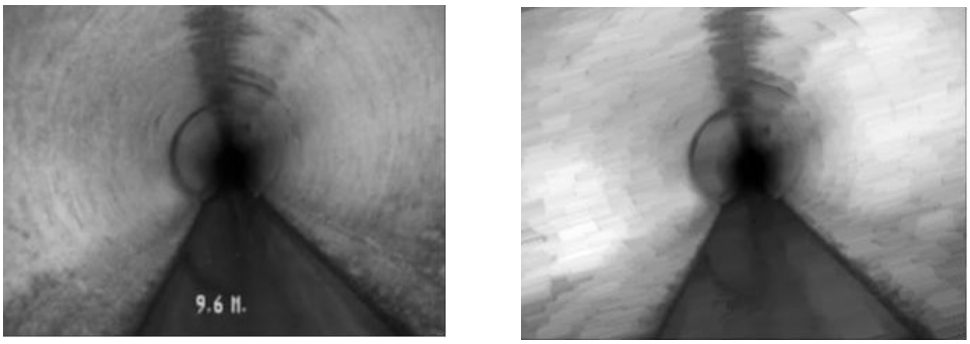


Figure 2-17. Example of Opening operation

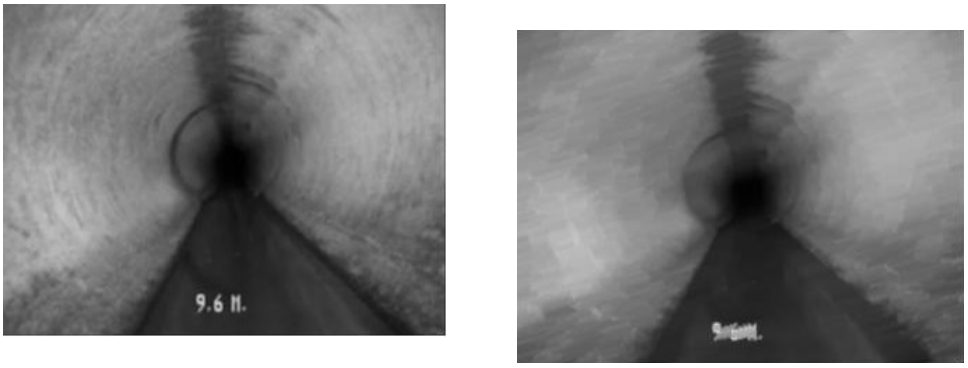


Figure 2-18. Example of closing operation

2.4.3 Image segmentation

Image segmentation is one of the most important steps in image analysis. It subdivides an image into its constituent regions or subjects. Considering the problem in hand, the subdivision level and its details would be defined (Gonzalez and Woods 2006). However, image noise and artifacts hinder segmentation operation from being done properly. Segmentation methods can be divided into three groups according to the dominant features they employ: Pixel-based, edge-based, and region-based. Pixel-based methods only use the gray values of the individual pixels. Region-based methods analyze the gray values in larger areas. Finally, edge-based methods detect edges and then try to follow them in the image areas (Jahne 2002). In this section, these methods are illustrated shortly.

Pixel-based (point-based) segmentation is the simplest method in image segmentation. The connectivity among components is grouped based on pixel connectivity in which each pixel is labeled with the gray level of its group. Considering these intensity values, a brightness constant or threshold can be determined to segment objects and the background. As the oldest segmentation method, thresholding is computationally inexpensive and is still widely used in simple applications (Gonzalez and Woods 2006).

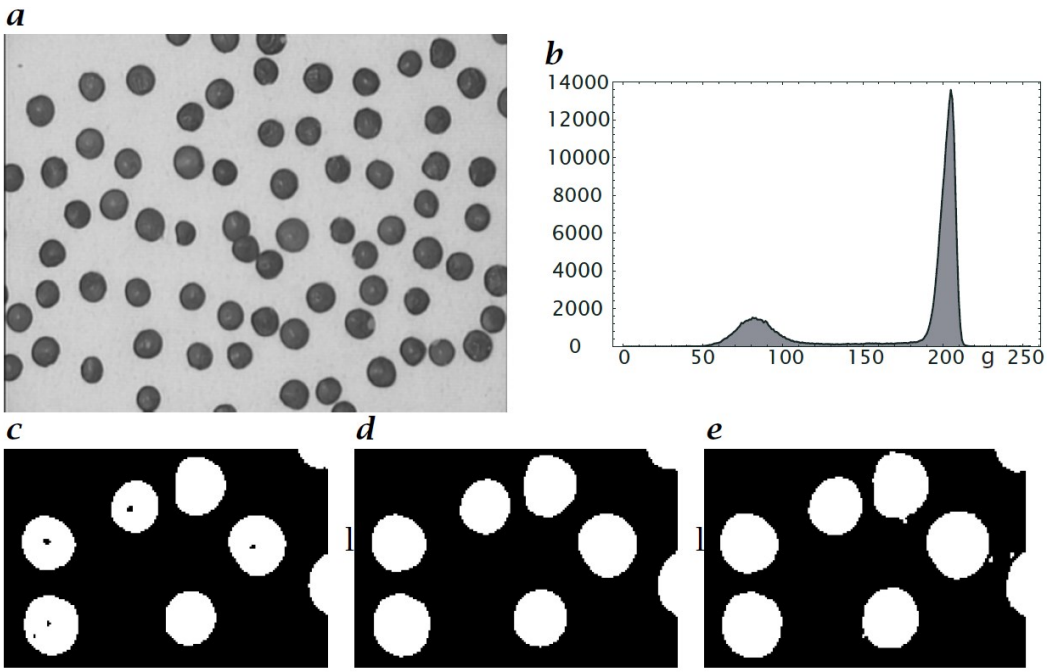


Figure 2-19. Pixel-based segmentation

(a) Original image; (b) image histogram; (c-e) segmentation with various global thresholds. (Adapted from (Jahne 2002))

The common limitation of pixel-based segmentation is that it does not take into account the local neighborhood. Also, when there is not a constant gray value in the objects, the size of the segmented objects cannot be determined accurately (Jahne 2002).

Edge-based segmentation detects and links edge pixels to form corners. It usually works with less complex algorithms comparing region-based methods. The remarked edges define highlight locations of discontinuities in gray level, color, texture, etc. The output image will be processed by a series of sequential algorithms to result in continuous edges through segmented parts (Jahne 2002). However, this process can be influenced by image noise, and the algorithm cannot clearly detect edges.

Region-based segmentation aims to detect regions in an image directly instead of defining borders then separating the regions as done in other algorithms. Introduced algorithms, try to classify pixels based on their grey values and neglect the integrity of the object. There are common techniques in region segmentation methods such as region growing, split and merge, watershed segmentation. The explanation of these techniques is out of the scope of this research and for further study please refer to Jain (1989), Gonzalez and Woods (2006) and Sonka et al. (2007).

Analyzing the image regions is a critical step in image classification and recognition. Image features can be extracted as a collection of data and be quantized in a feature vector for the analyzed image.

2.5 Machine Learning

Machine learning is a subgroup of artificial intelligence in which computer algorithms are trained to recognize the patterns of data and decide in new data automatically. Machine learning methods have been widely used in both feature extraction and defect classification. The algorithms can learn from data with or without human intervention. There are different training algorithms, and each can be employed, considering the type of problem in hand. The main learning mechanisms are supervised and unsupervised (or self-organized). A brief explanation of each one of the mentioned mechanisms is provided in the following.

Supervised learning

The supervised learning mechanism trains the algorithm by a set of labeled input data. A dataset is provided with instances of input stimuli and corresponding target values. The network applies the calculated weights and represents the outputs. The output results are continuously compared with the desired ones. Using a learning rule error between the actual output and the desired output is calculated to adjust the network's weights. Therefore, after several iterations, the actual output becomes the closest match to the target output.

Unsupervised learning

Unsupervised learning does not need supervision, and there is not a defined output to compare with. Unsupervised learning relies on presented input patterns and training data. The algorithm arbitrary discovers emergent collective properties and organizes the patterns into categories. Generally, the problem of pattern recognition and defect classification from sewer pipeline images is typically ill-posed since the proposed models are not able to be generalized for unseen images.

2.6 Convolutional Neural Network

The idea of Convolutional Neural Networks (CNNs) for image recognition was first developed by Fukushima (1975) under the name of the Neocognitron. Later in 1998, LeCun, Bottou, Bengio, and Haffner introduced LeNet-5 (Lecun et al. 1998). CNNs are basically conventional neural networks with two extra layers as convolutional layers and pooling layers in addition to fully connected layers and activation functions. CNN is able to assume the images with any size as an explicit input. A rectangular receptive field slides across over dimensions of the image to calculate kernel weights that are shared in all slides of the layer. In result, the model achieves a remarkable reduction in the number of parameters and calculations in the network. Therefore, the local space of the features will not be important to any further extent. In CNNs the pre-processing and feature definition and extraction steps are omitted, so feature extraction and classification both are included in one single structure (Figure 2-20).

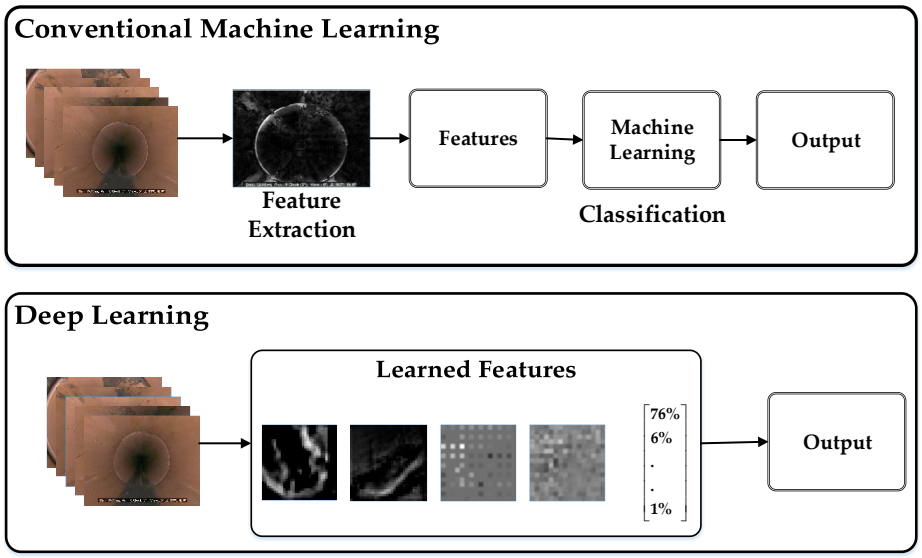


Figure 2-20. Machine learning approach and deep learning approach comparison (Moradi et al. 2019a)

2.6.1 CNN architecture

A typical CNN includes a set of layers that each layer contains one or more planes. The architecture put together a series of convolutional and pooling layers after the input layer and ends with fully connected layers in final layers to present the model's prediction. The planes can be employed as feature maps, and each plane has a related feature detector in a local receptive field which is sliding across the planes in the previous layer. The input image passes through the various convolutional and pooling layers and gets smaller and meanwhile richer features are extracted by convolutional layers. So, the primary layers extract low level features and the next layers interact with higher levels and determine spatial combinations of the lower features in the previous layers. Moreover, the network is trained with the usual backpropagation gradient-descent procedure.

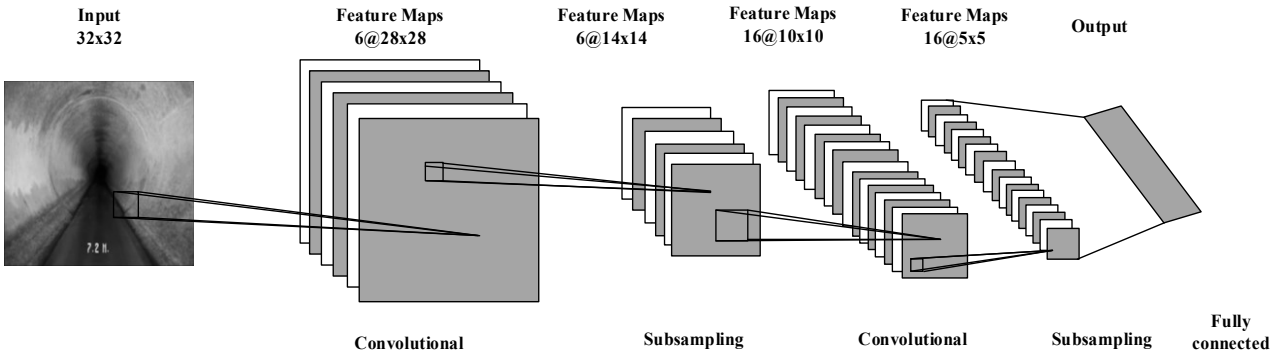


Figure 2-21. A typical scheme of Convolutional Neural Network

2.6.2 Convolutional layers

The pixel values of the input image feed to convolutional layers in which neurons act as filters. The input size of neurons in convolutional layers is the size of the receptive field, which is sliding through input image dimensions. The output of filters in each layer is a feature map and feeds to the next convolutional layer as an input. The filters survey across the whole previous layer with moving one *stride* each step. The receptive fields have overlap by *field width – stride*. In case the previous layer size is not dividable by the size of the receptive field, then the filter will miss the information in the edges of the input feature map. To solve this problem, zero padding can be employed by adding zeros to the edges of the input.

2.6.3 Pooling layer

Another extra layer in CNNs comparing to conventional ANNs is pooling layer. The pooling layers decrease the number of parameters in the network by downsampling the extracted features in the previous layers. The down sampling function increases the computational speed and also prevents the network from overfitting. In pooling layers, like convolutional layers a receptive field slides through the extracted feature maps by a predefined stride (*s*). Generally, the size of stride and receptive fields in pooling layers are considered equal, so there will not be any overlapping among them. There are different types of pooling methods such as Max pooling, Average pooling, and L2-norm pooling. Figure 2-22 shows an example of max pooling function in a pooling layer.

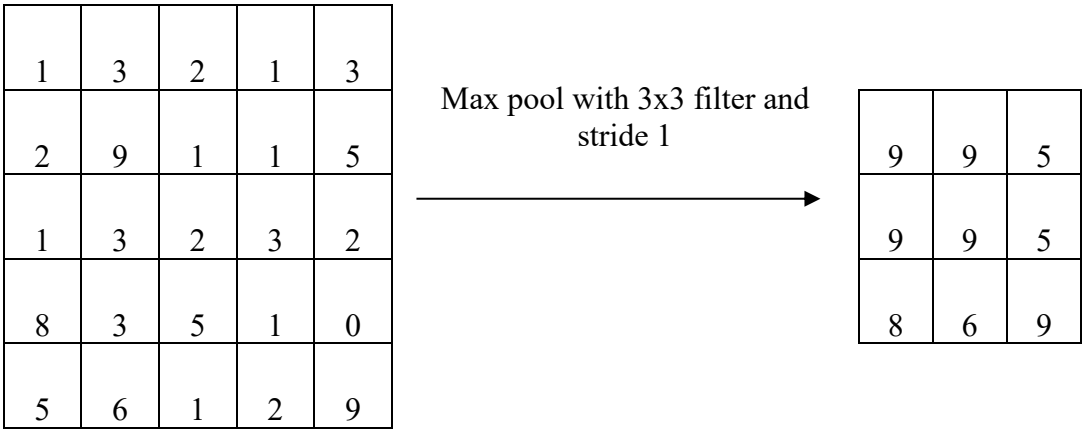


Figure 2-22. Max pooling function with 2 x 2 filter size and stride 1.

2.6.4 Fully connected layer

Fully connected layers stack on top of the network to flat the feature maps into feature vectors and finally predicting the class probabilities using activation functions like softmax. In a classification problem, the output would be a vector with the length of the number of classes, and each number in the classes indicates the probability of a certain class. For example, in the vector of 10 classes [0.2, 0, 0, 0, 0, 0, 0.1, 0.6, 0, 0.1] there is a probability of 20% that image be in class 1, 10% be in class 7, 60% be in class 8, and 10% be in class 10.

2.6.5 Activation functions

Activation functions empower the neural networks with nonlinearities which is the main difference of neural networks and linear regressions. The main application of activation functions is to determine upon receiving the information if a neuron should be activated or should ignore it. Technically, activation function provides nonlinearity over the input signals, so in backpropagation the gradients can be supplied with error to update the weights and losses in the network (Equation 2-1).

$$Y = Act.(\sum(w_i * n_i) + b) \qquad \textbf{Equation 2-1}$$

Where *Y* is the activation function, w_i is weight, and n is the neuron, b is the bias.

Sigmoid

Sigmoid activation function is widely used in neural networks since it is differentiable and imposes nonlinearity (Equation 2-2).

$$Y = 1/(1 + e\text{^} - x) \qquad \text{Equation 2-2}$$

As shown in figure 2-23, it ranges from 1 to 0 with an *S* shape.

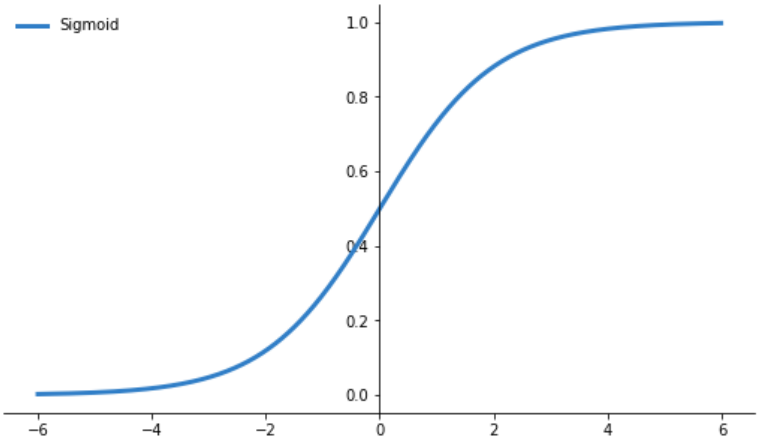


Figure 2-23. Sigmoid activation function

The main application of sigmoid activation function is the values classification. The main problems with sigmoid activation function are that it saturates beyond -3 and +3, so in those regions gradients become too small and resembling zero and results in the network stops learning. Moreover, the values in the function only range from 0 to 1, so all the received values are positive.

Tanh

Tanh function maps the neuron values in ranges from -1 to 1(Equation 2-3).

$$tanh(x) = 2/(1 + e\text{^}(-2x)) - 1 \qquad \text{Equation 2-3}$$

As shown in figure 2-24 tanh is continuous and differentiable over all the values. Also, it is nonlinear and in backpropagation, the errors can be easily considered.

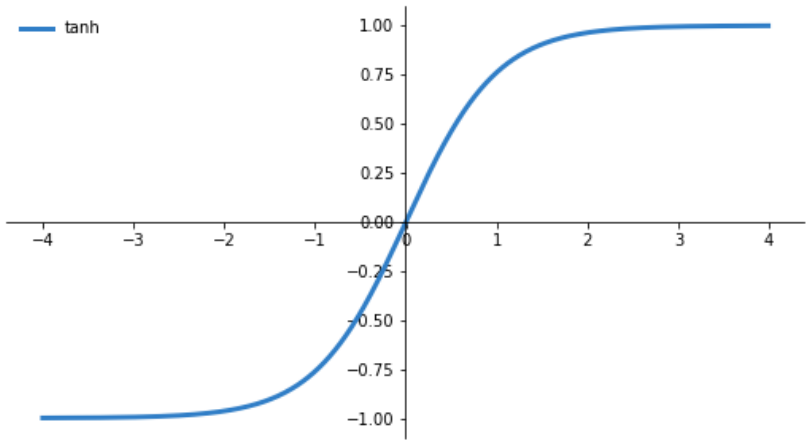


Figure 2-24. Tanh activation function

The main application of tanh activation function is when it is required to classify a class with higher gradient values. The problem with tanh function is vanishing gradients where the function becomes flat at -3 and 3 regions.

ReLU

Rectified linear unit (ReLU) is the most widely used activation function especially in deep learning algorithms since it does not saturate on positive values and also it is fast and has low computational complexity (Equation 2-4).

$$Y = max(0, x) \qquad \text{Equation 2-4}$$

ReLU function is nonlinear, so backpropagation is possible. It does not activate all neurons at the same time, and negative values will be converted to zero. Figure 2-25 shows how ReLU activation function works.

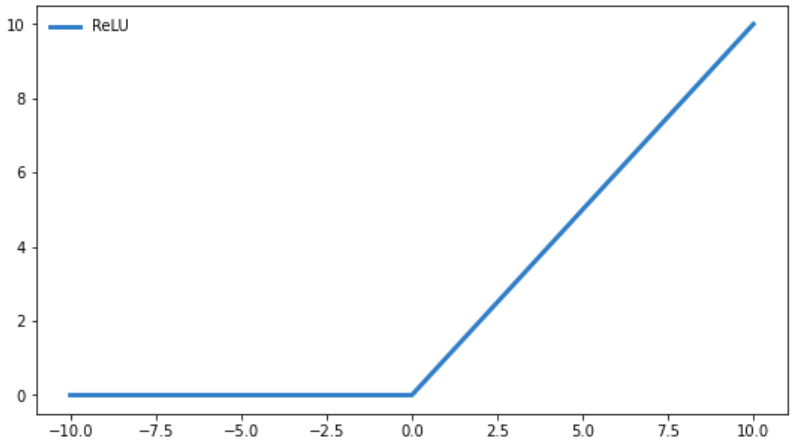


Figure 2-25. ReLU activation function

The main disadvantage of ReLU activation function is known as dying ReLUs, which means during training, some neurons are killed, and they only output zero so they will not in training weight updates anymore.

Leaky ReLU

This function is an improved ReLU function since in ReLU for negative values, the gradient is zero, and neurons are deactivated in that region. Leaky ReLU is introduced to solve this problem. In this function, a small linear hyper-parameter is defined for negative values (Equation 2-5).

$$f(x) = \begin{cases} ax, x < 0 \\ x, x >= 0 \end{cases} \qquad \text{Equation 2-5}$$

The hyper parameter α defines the amount of leak in the function and ensures that neurons never die (Figure 2-26).

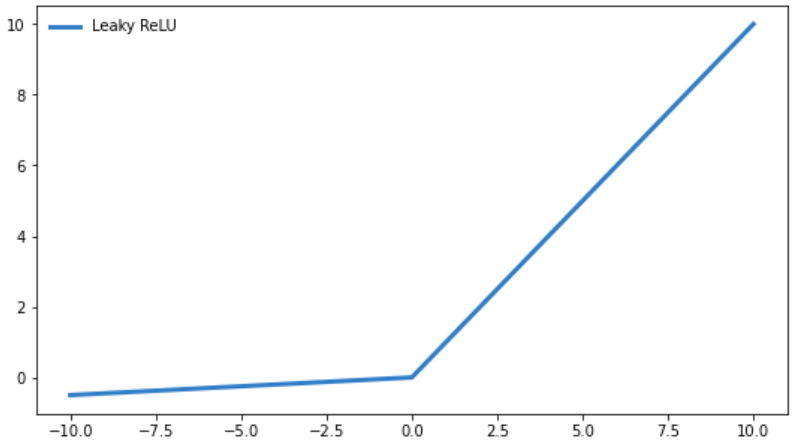


Figure 2-26. Leaky ReLU activation function

Softmax

This function is acting like sigmoid activation function and applicable in classification problems. In contrary to sigmoid function that is only able to handle two classes, softmax can conduct multiple classes. Softmax activation function considers the output of each class between 0 and 1 and then divides by the sum of the outputs (Equation 2-6).

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \qquad \text{Equation 2-6}$$

The softmax function usually is used in the last layer to predict the probability of the classes.

2.6.6 CNN architectures

A typical CNN architecture stack a couple of convolutional layers followed by a pooling layer, then a more set of convolutional and pooling layers. After each convolutional layer, an activation layer is placed. In recent years, many improvements have been achieved in increasing the accuracy of the CNNs. State-of-the-art architectures now can achieve a lower error rate near human vision. Competitions as ILSVRC ImageNet (Russakovsky et al. 2015) each year introduce top image classifiers. In this section, the top classifiers are introduced through their proposed architecture.

LeNet-5

LeNet-5 developed by LeCun et al. (1998) is one of the most popular CNN architectures. It is commonly used in natural language processing (NLP) tasks like handwritten digits recognition. LeNet-5 consists of 7 layers including three convolutional layers (C1, C3, and C5), two pooling layers (S2 and S4), and one fully connected layer (F6) and followed by output layer. The kernel size for convolutional layers is 5×5 with stride one and for pooling layers is 2×2. Input images from MNIST dataset are zero-padded from 28×28 to 32×32 pixels and the size of images decreases through the network. Table 2-2 represents the architecture of LeNet-5.

Table 2-2. LeNet-5 architecture

Layer	Feature Map	Size	Kernel Size	Stride	Activation
FC	-	10	-	-	softmax
FC	-	84	-	-	tanh
Conv2d	120	1x1	5x5	1	tanh
Average Pooling	16	5x5	2x2	2	tanh
Conv2d	16	10x10	5x5	1	tanh
Average Pooling	6	14x14	2x2	2	tanh
Conv2d	6	28x28	5x5	1	tanh
Input	1	32x32	-	-	-

AlexNet

AlexNet network is very similar to LeNet-5 but much deeper and includes more parameters. It was developed by Krizhevsky et al. (Krizhevsky et al. 2012) and won the 2012 ImageNet ILSVRC challenge. The convolutional layers stack on top of other convolutional layers in the last set of convolutional layers. To avoid overfitting, authors applied dropout method (0.5 dropout rate) in training. Moreover, data was augmented by various transformations such as flipping and offsetting. Figure 2-27 shows the architecture of AlexNet.

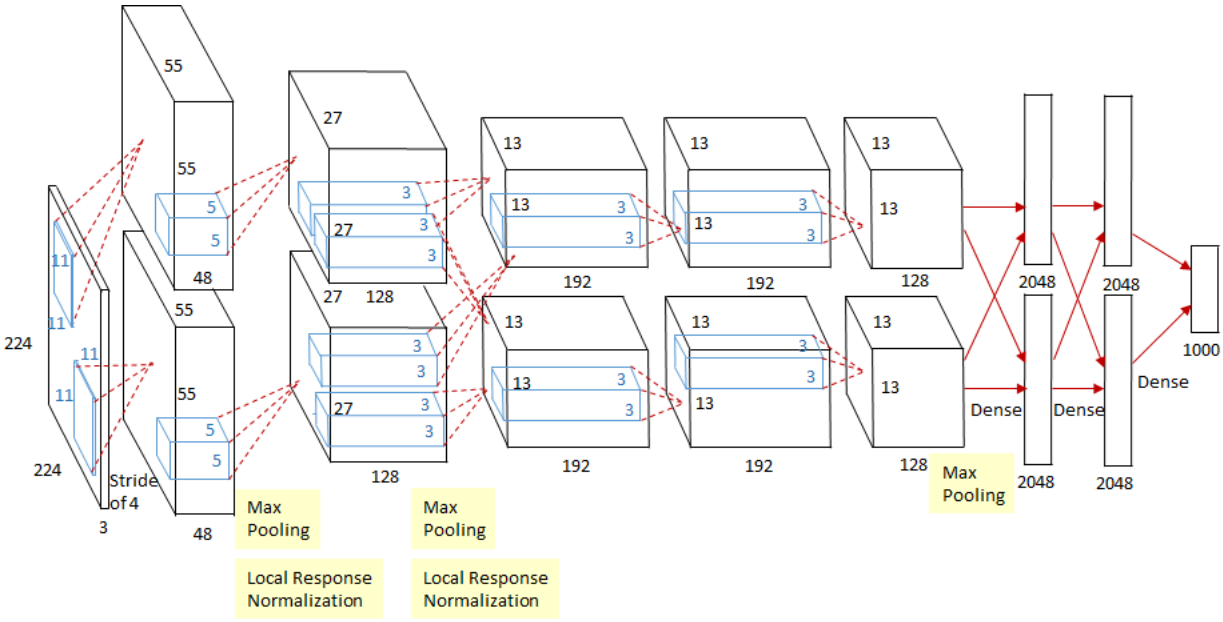


Figure 2-27. AlexNet architecture (adapted from (Krizhevsky et al. 2012))

The network was split into two sections to train simultaneously on different cores. AlexNet employed local response normalization after ReLU step, which makes the neurons that most strongly activate inhibit neurons at the same location (Krizhevsky et al. 2012). This results in more generalization capacity of the model since the network searches for wider variety of features by pushing the features apart. The layers of AlexNet architecture are described in Table 2-3.

Table 2-3. AlexNet architecture

Layer (type)	Output Shape	Param #	Activation
input_1 (InputLayer)	(None, 227, 227, 3)	0	-
conv2d_14 (Conv2D)	(None, 55, 55, 96)	34944	ReLU
max_pooling2d_9	(None, 27, 27, 96)	0	-
conv2d_15 (Conv2D)	(None, 27, 27, 256)	614656	ReLU
max_pooling2d_10	(None, 13, 13, 256)	0	-
conv2d_16 (Conv2D)	(None, 13, 13, 384)	885120	ReLU
conv2d_17 (Conv2D)	(None, 13, 13, 256)	884992	ReLU
max_pooling2d_11	(None, 6, 6, 256)	0	-
flatten_4 (Flatten)	(None, 9216)	0	-
dense_9 (Dense)	(None, 4096)	37752832	ReLU
dropout_5 (Dropout)	(None, 4096)	0	-

dense_10 (Dense)	(None, 4096)	16781312	ReLU
dropout_6 (Dropout)	(None, 4096)	0	-
dense_11 (Dense)	(None, 1000)	4097000	Softmax
Total params: 61,050,856			
Trainable params: 61,050,856			
Non-trainable params: 0			

VGGNet

VGGNet was developed by Simonyan and Zisserman (2014a) and was one of the top competitors in ILSVRC 2014. VGGNet includes 16 convolutional layers and presents a deep and uniform architecture. VGGNet is used as a preferred network for image feature extraction. The model configuration and weights are available publicly to be used in many other applications. However, the high number of parameters (i.e., 138 million) makes it a bit challenging to handle.

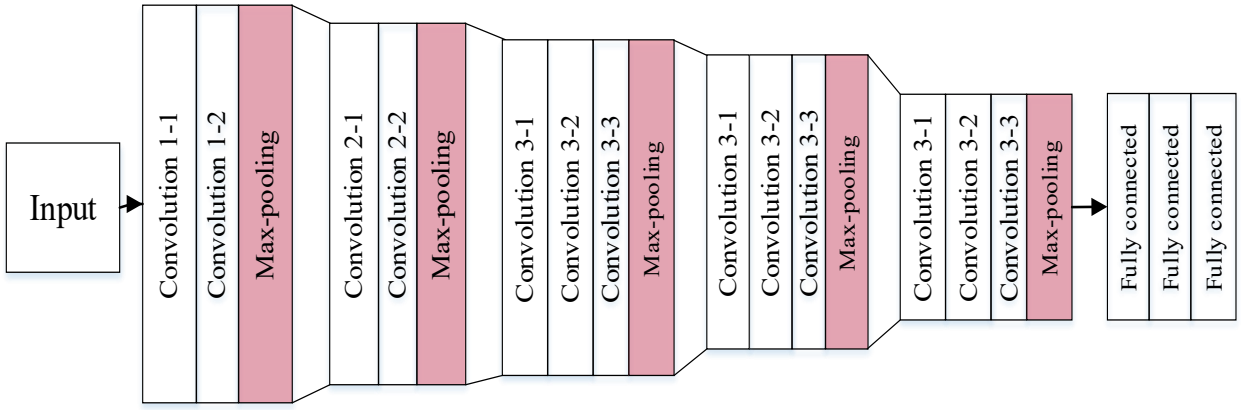


Figure 2-28. VGGNet architecture

GoogLeNet

Google developers Christian Szegedy et al. (2014) developed GoogLeNet/Inception which was the winner in the ILSVRC 2014 competition. The error rate was 6.67% which was very impressive and almost near human level accuracy. This was achieved by employing sub-networks called inception modules which resulted in much fewer parameters comparing previous networks like AlexNet (around 6 million instead of 60 million parameters). The 22-layer deep CNN utilized techniques like batch normalization, image distortions, and RMSprop.

Figure 2-29 shows the architecture of inception module. The convolutional layers use various kernel sizes (1×1, 3×3, 5×5) making them able to capture the features at different scales. All convolutional layers use ReLU as the activation function. Moreover, the output is the same size as input since all the layers use the same padding with stride 1. This results in concatenate all the layers outputs along the depth dimension (Szegedy et al. 2014).

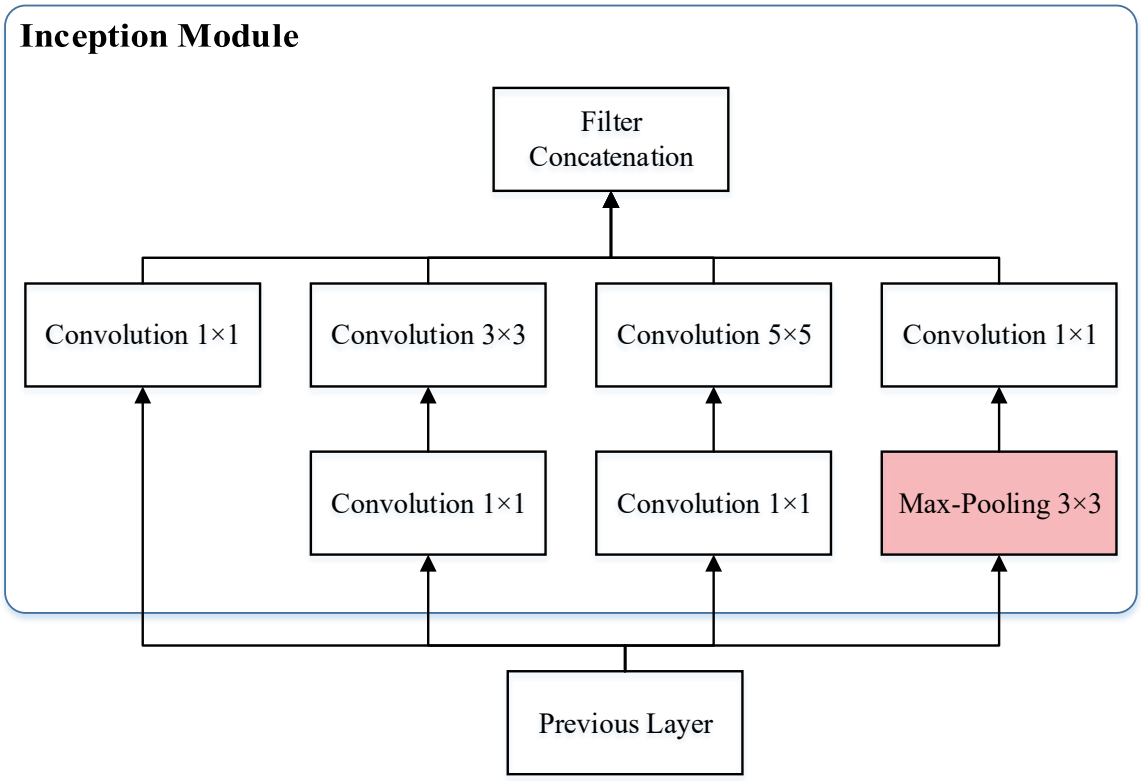


Figure 2-29. Inception module

ResNet

Residual Neural Network (ResNet) introduced by Kaiming et al. (2015) at the ILSVRC 2015. The authors proposed gated units called skip connections similar to elements applied in recurrent neural networks (RNN) which make them able to train a network with 152 layers and error rate of 3.57% that is better than human level accuracy. So, the signal feeding as an input to a layer is also added to the output of a layer on a higher stack (Kaiming et al. 2015). ResNet architecture is simple, and the network starts and ends the same as GoogLeNet. Each residual module consists of two convolutional layers with ReLU activation function. The kernel size is 3×3, and like inception, modules keep the output size equal to the input size.

2.7 Object detection models

In addition to defect classification, the location of the defect in the pipe is important since in many sewer assessment protocols, the detected defect position is considered in severity evaluation. Generally, the assignment of detected instances in an image to a certain class is called object detection. In recent years, object detection algorithms evolve from image processing techniques which require complex feature engineering and numerous mathematical operations such as Viola-Jones object detection framework (Viola and Jones 2001), to deep neural networks that represent capabilities to perform real time object detection with acceptable accuracies. In following, the most popular object detection architectures are introduced and explained briefly.

2.7.1 R-CNN

The region-based convolutional neural network (R-CNN) (Girshick et al. 2014) proposes to check for objects in region proposals instead of a massive number of regions. R-CNN employs an external algorithm called selective search to extract the regions. The selective search identifies different regions in an image based on capturing all object scales and diversifications (Uijlings et al. 2013). The warped region proposals are fed forward to a trained CNN model to get the region of interest for each image, and then a support vector machine (SVM) classifies objects and background for the regions. Finally, using a linear regressor the objects bounding boxes are generated in the classified images.

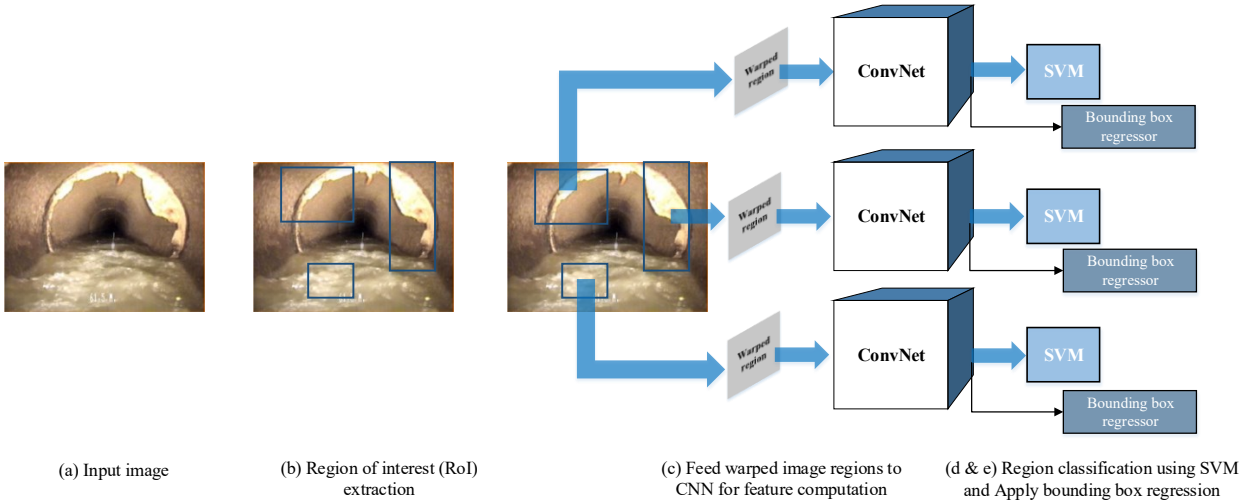


Figure 2-30. R-CNN typical architecture

The multistage training pipeline of R-CNN makes it computationally expensive and time-consuming. Also, each region proposal should pass three models for feature extraction, classification, and regression, so the prediction would be relatively slow, especially when the model deals with large datasets. Moreover, since there are FC layers in the CNN model, the input sizes must be fixed, and the algorithm proposes warp or crop region proposals. This may impose re-computation for each region and also due to the warping operation, the object can be placed partly in a cropped region and leads to a reduction in recognition accuracy.

2.7.2 Fast R-CNN

Fast R-CNN was introduced by Girshick (2015) to improve the R-CNN detection speed. In this approach, instead of performing the CNN for each region proposals, the convolution process is conducted for each image and all regions of interest (Girshick 2015). The author proposed an architecture to generate convolutional feature maps by processing the entire image with several convolutional and pooling layers and then convert each region proposal into a feature vector (Girshick 2015). The extracted feature vectors are fed to the fully connected layers for classification using softmax probability estimation of the $C+1$ classes (C object classes and one background class) and regression for encoding bounding box positions with four real numbers for predicted classes. Thereby, Fast R-CNN uses a single model instead of three different models in

R-CNN and Fast R-CNN training time would be much faster. The detection required time is also reduced while the accuracy is improved by the method.

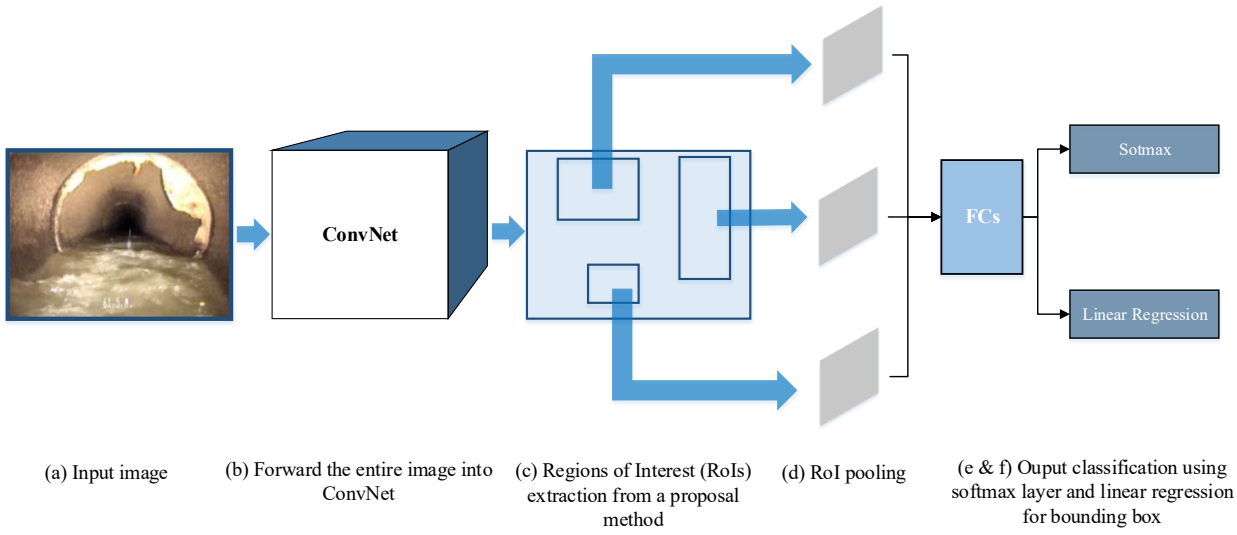


Figure 2-31. Fast R-CNN architecture

2.7.3 Faster R-CNN

The fast R-CNN has certain limitations as it uses selective search as a proposed method to identify Regions of Interest, which still is slow and time-consuming when dealing with large datasets. To overcome these limitations, the faster R-CNN developed by Ren et al. (2017) uses Region Proposal Network (RPN) that identifies region proposals and shares the convolutional features with the detection network. RPN creates an optimized set of region proposals with higher quality compared to those generated by the selective search method. Finally, RPN and Fast R-CNN are merged into a single network while sharing their convolutional features, then the unified network can classify and output the bounding boxes for objects (Ren et al. 2017).

The idea of anchor boxes is introduced by authors to cope with the differences in the objects aspect ratio and scale in the images. At each location, three types of anchor boxes for scale 128×128, 256×256, and 512×512 are used (Ren et al. 2017). In the same way, for aspect ratio, the model applies three aspect ratios 1:1, 2:1, and 1:2. Thereby, for each location, nine boxes are presented, and RPN predicts if the box is background or foreground object (Ren et al. 2017). The Faster R-CNN, like the other discussed object detection frameworks, does not capture the whole image regions at once and to detect the objects it requires a sequence of passes.

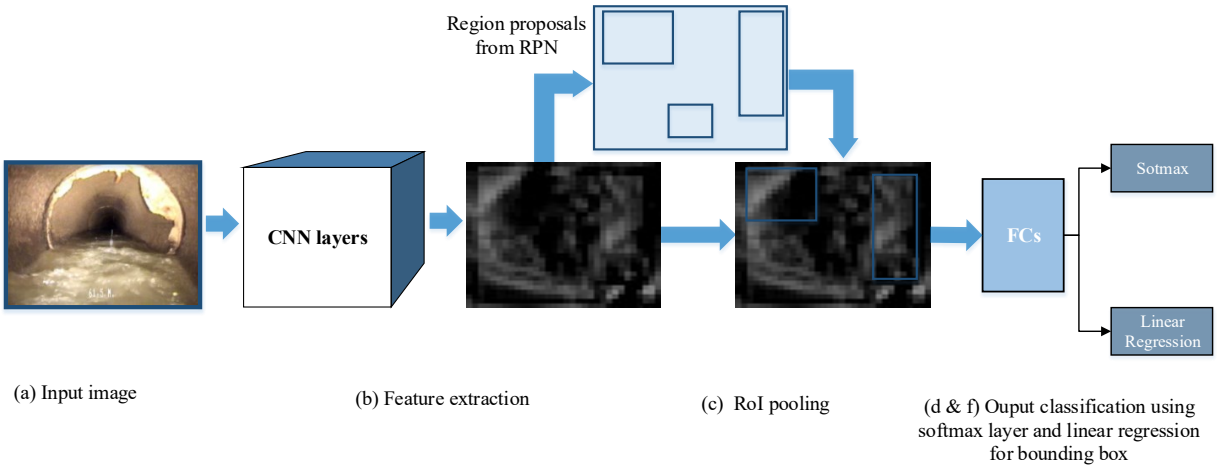


Figure 2-32- Faster R-CNN architecture

All the introduced approaches mainly rely on a sliding window to train the classifier for object detection among a wide range of proposal boxes within various size scales and positions. In some models such as Faster R-CNN the weights can be shared; however, the main computational complexity that derives from convolving filters with the whole image remains especially in large image inputs (Mnih et al. 2014).

2.7.4 YOLO

Redmon et al. (2016) introduced a novel framework called You Only Look Once (YOLO), which directly predicts both confidences for class probabilities and bounding boxes in a single evaluation. YOLO divides the input image into an S×S grid, and each grid cell predicting the objects bounding boxes and their respective confidence scores (Redmon et al. 2016). This confidence is simply indicating the probability of objects existence ($Pr(Object) \geq 0$) and represents the confidence of the prediction (IOU_{pred}^{truth}). Also, in each grid cell C condition of class probabilities ($Pr(Class_i|Object)$) needs to be predicted, so for each individual box, class specific confidence scores are calculated (Equation 2-7) (Redmon et al. 2016):

$$Pr(Object) * IOU_{pred}^{truth} * (Pr(Classi|Object) = \Pr(Class_i) * IOU_{pred}^{truth} \quad \text{Equation 2-7}$$

In the proposed YOLO framework, 24 Conv layers and 2 FC layers inspired by GoogLeNet is used. 1×1 reduction layer followed by 3×3 Conv layers replace initial inception modules. In the final layer for every grid cell, a prediction tensor is generated to define the estimated probabilities for each class, the number and coordinates of anchor boxes, and a confidence value (Redmon et al. 2016). With all improvements in speed and detection accuracy, YOLO still has a limitation in detecting small objects due to spatial constraints caused by bounding box predictions (Redmon et al. 2016). Moreover, YOLO is not able to generalize to objects with unseen aspect ratios and new construction (Zhao et al. 2019).

2.8 Conventional visual inspection and assessment

Sewer pipeline inspection using mobile CCTV is the typical approach in visual inspection methods. CCTV system uses a television camera mounted on a robot which is in conjunction with a display monitor and a recording device. The robotic system is placed in the pipeline and directs

by a trained operator through the pipe. Human operators are required to be trained and be familiar with the protocols and pipeline grading system. During the operation, the operator has to stop the robot to focus on suspicious areas to closer inspection and collect more evidence to decide on detection, classification, and criticality of the existing defects against documented protocols. Therefore, there would be a considerable amount of stops and starts, which increases the time of inspection and make it costly as well. Also, the videos might be checked off-site by specialized operators to either recheck the reports or inspection. The existing physical condition of the pipeline can then be assessed using mentioned standard protocols as PACP, WRc, or Municipalities condition evaluation guidelines. This results in the identification of the deterioration pattern and determination of the potential collapse or failure of an asset (Rahman and Vanier 2004).

Both in site and offline reporting procedures are highly dependent on the operator's skill and processing the data provided by CCTV is time-consuming and labor-intensive. There is a probability of missing some of the defects that are hidden from the camera by obstructions and certain types of defects that cannot be captured by CCTV such as those are under waterline. Moreover, operator skillfulness, fatigue, and concentration may affect the reliability and consistency of inspection results.

2.9 Automated defect detection and condition assessment

CCTV inspection reports have some limitations such as lack of geometric references, subjective assessment based on the operator's skill, and image quality variation. Recent developments in digital imaging industry have led to a remarkable cutback in the cost of visual inspections for municipalities. In many assessment protocols, it is required to provide visual evidence for detected defects or faults supporting condition assessment of the pipeline. Recorded images from the pipe offer a complete set of data such as features, patterns, position, and severity of the faults. Moreover, the latest advances in processing capacity of the computers and accessibility of cloud computing paved the way for the employment of powerful machine learning algorithms and computer vision techniques. In the last decade, state-of-the-art artificial intelligence algorithms have been suggested by the studies in the field to automate the sewer pipe inspection and assessment.

The research works generally analyze the inspection videos using image processing algorithms like morphological operations and image segmentation, then employing a machine learning algorithm for defect classification. However, in recent studies both feature sampling and classification are carried out deep neural networks. Regarding the utilized computer vision technique, the studies in sewer defect detection automation are categorized into three groups: morphology, feature extraction, and detection/recognition.

Studies in the morphology group extracted sewer defects features using morphological operations and proposed a defect detection model based on geometrical features of the defects. Research works in feature extraction utilized various image processing techniques to process the sewer images regions and detect the defects based on extracted features using a machine learning algorithm. In the third group, the models proposed an approach to conduct feature extraction and defect detection and classification in one framework simultaneously. Figure 2-23 shows different categories of research works in sewer defect detection automation.

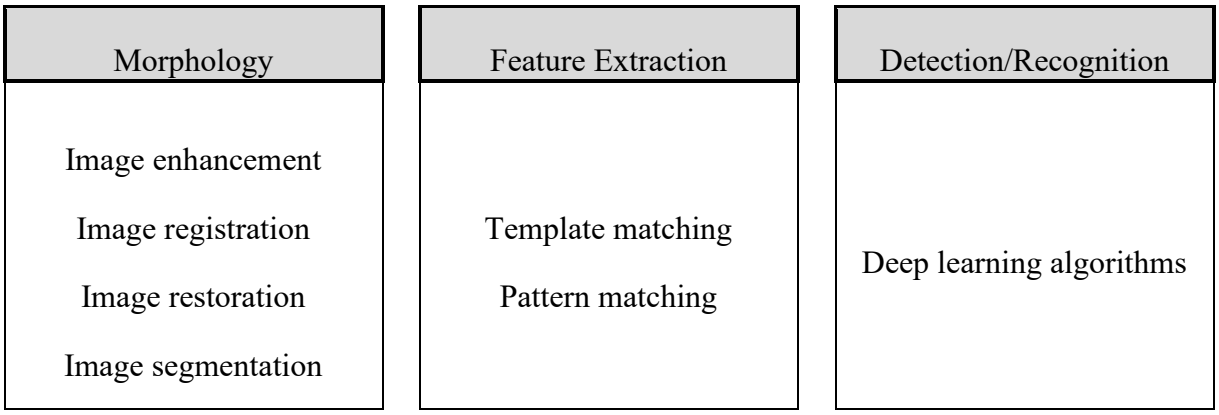


Figure 2-33. Computer vision techniques used in sewer defect detection (adapted from (Moradi et al. 2019a))

2.9.1 Morphology

Morphological operators employ various non-linear mathematical operations to describe the structuring elements of image features. Morphological operators can establish the pixel values in the image considering the order quantified by the operator and comparing the equivalent pixel in the image with the pixel neighborhood.

Morphological operations have been a common tool for sewer image processing in previous research. Researchers have employed a wide range of morphological operations to segment pipe defects accurately. A number of papers utilized a series of binary or greyscale segmentation techniques for edge detection. The studies differentiated the sewer defects from pipe walls to extract the geometrical features by edge detector operators and thresholding the histogram of segmented pixels (Chae et al. 2003; Chae and Abraham 2001; Guo et al. 2007, 2009a; b; Halfawy and Hengmeechai 2014a; Hawari et al. 2018; Iyer and Sinha 2005, 2006; Kirstein et al. 2012; McKim and Sinha 1999; Moselhi and Shehab-Eldeen 1999; Pan et al. 1994; Shehab and Moselhi 2005; Sinha et al. 2003; Sinha and Fieguth 2006b; a; Sinha and Knight 2004; Sinha Sunil K. 2001; del Solar and Köppen 1996; Yang and Su 2009). Also, morphological segmentation based on edge detection (MSED) is used to identify the morphology representatives for sewer pipe defects on CCTV images such as cracks, joints, and holes (Dang et al. 2018; Su et al. 2011; Su and Yang 2014) and MSER algorithm for text detection in sewer images (Dang et al. 2018). Table 2-4 shows the studies using morphological operations.

The patterns in the sewer pipe images can be interpreted by dark and light binary shapes; therefore, morphological operations are applicable tools for analyzing the images (Koch et al. 2015). Algorithms like edge detection and thresholding are suitable to capture the shapes and are applicable to segment some defects like cracks. However, morphology algorithms are highly dependent on image quality. The conditions of the recorded images from the pipe such as illumination, noise, and low contrast may affect the segmentation and consequently defect detection accuracy.

Table 2-4. Studies in automated sewer defect detection using morphological operation

Author(s)	Year	Data Acquisition	Pipe Type	Method	Feature Descriptor	Defect(s)	Classifier	Assessment Protocol
Hawari et al.	2018	CCTV	VCP	Gabor Filters	Geometrical			PACP
Dang et al.	2018	CCTV	Concrete	MSER, MFI	-	Crack	OCR	-
Halfawy, & Hengmeechai	2014	CCTV	VCP	Edge Detection, Hough Transform	-	Crack	RuleBased	WRc
Su & Yang	2014	CCTV	VCP	MSED, OTHO	-	Crack, Joint	-	-
Kirstein et al.	2012	CCTV	VCP, Concrete	Edge Detection, Hough Transforms	-	Flow-line	RuleBased	-
Su et al.	2011	CCTV	VCP	MSED	Geometrical	Crack, Break, Debris, Joint	-	-
Guo et al.	2009	RedZone	VCP, Concrete	Edge Detection	Gradient Based	ROI	Rule Based	PWSA
Gou et al.	2009	CCTV	VCP	Histogram Matching	-	ROI	Change Detection	PACP
Yang & Su	2009	CCTV	VCP	Otsu	-	Broken Pipe, Crack, Fracture, And Open Joint	RBN	-
Gou et al.	2007	RedZone	VCP	Edge Detection	-	ROI	-	PACP

Author(s)	Year	Data Acquisition	Pipe Type	Method	Feature Descriptor	Defect(s)	Classifier	Assessment Protocol
Sinha & Feiguth	2006	SSET	Concrete	Edge Detection, Otsu	Geometrical	Crack	Rule based	-
Sinha & Feiguth	2006	SSET	Concrete	Segmentation	Geometrical	Cracks, Holes, Joints, Laterals, Collapse	Rule Based	-
Iyer & Sinha	2006	SSET	Concerete	Segmentation	Geometrical	Crack	Rule Based	NAAPI
Sinha & Feiguth	2006	SSET	Concrete	Segmentation	Geometrical	Crack, Joint, Lateral	Rule Based	-
Shehab & Moselhi	2005	CCTV	VCP	Segmentation	Geometrical	Infilteration	ANN	-
Iyer & Sinha	2005	SSET	Conceret	Segmentation	Goemetrcal	Crack	Rule Based	-
Sinha	2004	SSET	Conceret	Segmentation	Goemetrcal	Crack	ANN	CATT
Sinha et al.	2003	PSET	Concrete	Segmentation	Geometrical	Crack, Joint, Lateral	Rule Based	-
Chae et al.	2003	SSET	-	Segmentation	Geometrical	Crack, Joint, Lateral	ANN	-
Chae & Abraham	2001	SSET	Concrete	Segmentation	Geometrical	Crack, Joint, Lateral	Fuzzy-ANN	-
Sinha	2001	CCTV	Concrete	Segmentation	Geometrical	Crack		-

Author(s)	Year	Data Acquisition	Pipe Type	Method	Feature Descriptor	Defect(s)	Classifier	Assessment Protocol
Chae	2000	SSET	Concrete			Crack, Joint, Lateral	Fuzzy-ANN	-
McKim & Sinha	1999	SSET	Concrete	Segmentation	Geometrical	Crack, Joint, Lateral	Rule Based	-
Moselhi & Shehab	1999	CCTV	VCP	Segmentation	Geometrical	Crack	ANN	-
Solar & Koppen	1996	CCTV	-	Gabor, Segmentation	Geometrical	-	ANN	-
Pan et al.	1994	Image	-	Hough Transform	Geometrical	Joint	Rule Based	-

2.9.2 Feature Extraction

One of the main phases in image recognition is analyzing the image pixels and regions either locally or globally. Feature extraction creates an arrangement of distinguishable data and quantizing them in a numeric feature vector. Various contributions in design and employment of innovative feature extraction techniques including wavelet transform and co-occurrence matrices (Sinha et al. 1999; Sinha and Karray 2002; Ye et al. 2019), wavelet transforms and histograms of oriented gradients (HOG) (Halfawy and Hengmeechai 2014a; b; c; Wu et al. 2015; YANG et al. 2011; Yang and Su 2008; Ye et al. 2019), GIST descriptor (Myrans et al. 2018, 2019), and spatio-temporal features (Moradi et al. 2016; Moradi and Zayed 2017) have been made. Moradi et al. (2020) introduced an innovative method to extract sewer inspection videos using 3D-SIFT and classify the features by OC-SVM to identify anomalous frames. Some studies combined a series of feature extraction methods to come up with a unique feature vector (Fang et al. 2020; Mashford et al. 2010). One of the main approaches in studies is using image processing and computer vision techniques to determine geometrical features of the defects and train neural networks for defect classification (Chaki and Chattopadhyay 2010; Moselhi and Shehab-Eldeen 2000; Sinha and Fieguth 2006a).

Feature extraction methods are helpful to search space dimension reduction and reducing the number of calculations and computational cost. In sewer pipes because of the immense patterns and poor lighting conditions, it is almost impossible to define templates for defects. Cameras pose and illumination variation may reshape the defects forms and defects patterns can be confused easily. Thus, relying on the extracted features for defect classification is doubtful and the classifier gets confused because of indefinite features and patterns of sewer defects. Moreover, to train the classifier the features need to be engineered manually to label the data which is exhaustive and time-consuming. Table 2-5 shows the studies in automated sewer defect detection using feature extraction.

Table 2-5. Studies in automated sewer defect detection using feature extraction

Author(s)	Year	Data Acquisition	Pipe Type	Feature Descriptor	Defect(s)	Classifier	Assessment Protocol
Moradi et al.	2020	CCTV	VCP	3D-SIFT	ROI	OC-SVM	PACP
Xu et al.	2020	CCTV	PVC	HOG, LBP, Gabor, GLCM	ROI	Guassian-D	-
Ye et al.	2019	CCTV	Concrete, HDPE, PVC	Hu In variant Moments	Deformation, Collapse, Infiltration, Deposit, Joint	SVM	-
Myran et al.	2018	CCTV	VCP, Concrete, Brick	GIST	ROI	SVM	-
Myran et al.	2018	CCTV	VCP, PVC, Brick	GIST	Crack,Deformation, Collapse, Infiltration, Deposit, Joint	Random Forest	-
Moradi & Zayed	2017	CCTV	VCP	Spatio- Temporal	ROI	HMM	PACP

Author(s)	Year	Data Acquisition	Pipe Type	Feature Descriptor	Defect(s)	Classifier	Assessment Protocol
Wu et al.	2015	SSET	VCP	Haar Wavelet, Contourlet	Root, Collapse, Crack	AdaBoost, Random Forrest	-
Halfawy & Hengmeechai	2014	CCTV	VCF	HOG	Root	SVM	-
Mashford et al.	2014	Image		Haar Wavelet	Crack	SVM	-
Halfawy & Hengmeechai	2014	CCTV	VCF	HOG	ROI	Rule based	-
Yang et al.	2011	CCTV	VCp	Wavelet Transform	Joint, Crack, Break	Rule based	-
Mashford et al.	2010	PIRAT	Concerete	Gabor, HSB	Hole, Root, Joint, Deposit, Corrosion	SVM	-
Chaki & Chattopadhyay	2010	CCTV	VCP, Brick	Geometrical	Crack	Fuzzy-MDSS	-
Yang & SU	2008	CCTV	VCP, RCP	Wavelet Transform	Joint, Crack, Break, Fracture	SVM, ANN, RBN	WRc
Sinha & Fieguth	2006	SSET	Concerete	Geometrical	Joint, Crack, Break, Fracture	Fuzzy-ANN	-

Author(s)	Year	Data Acquisition	Pipe Type	Feature Descriptor	Defect(s)	Classifier	Assessment Protocol
Sinha & Karray	2002	PSET	Concerete	Geometrical, Co-occurrence Matrix	Crack	Fuzzy-ANN	-
Moselhi & Shehab	2000	Image	VCP	Geometrical	Crack, Joint, Deformation, Spalling	ANN	-
Sinha et al.	1999	PSET	Concerete	Fourier Transform, Karliunen-Loeve (KL) Transforin	Crack	Fuzzy-ANN	-

2.9.3 Detection and Recognition

In recent years there is a spike of using deep learning algorithms like Convolutional Neural Networks (CNNs). Accessibility to inspection data besides affordable computation machines made it practicable to utilize deep neural networks and improve the speed and accuracy of sewer assessment reports. Deep learning algorithms are used both in defect classification and detection in sewer pipelines. The main advantage of deep learning algorithms over the traditional machine learning algorithms is the capability to automatically extract the features during the training. Therefore, the crucial and time-consuming feature engineering stage used in typical machine learning algorithms is excluded. Deep neural network trains can get image as input data and spit out the defect class as the output. The first layers extract simpler attributes of the features, while deeper layers represent more complex features of the input image.

In recent researches, various architectures of convolutional neural networks (CNN) (Chen et al. 2018; Kumar et al. 2018; Li et al. 2019; Meijer et al. 2019; Moradi et al. 2018b; Xie et al. 2019), and fully connected networks (FCN) (Wang and Cheng 2019), and saliency model using recurrent neural network (RNN) (Wang and Cheng 2019) have been employed for sewer defect classification. Also, various object detection frameworks such as faster region-based convolutional neural networks (Faster-RCNN) (Cheng and Wang 2018; Kumar et al. 2020), and You Look Only Once (YOLO) (Kumar and Abraham n.d.; Yin et al. 2020), and Single Shot Detector (SSD) (Moradi et al. 2019c) have been used for sewer images defect detection. It is believed that CNNs have higher classification accuracy and better generalization capability comparing the other classification techniques (LeCun et al. 2015). Although application of deep learning algorithms seems to be the future trend in sewer defect detection automation, these algorithms need large training datasets. Meanwhile, the training process is computationally expensive especially for deeper networks with many layers. In recent years, the drawbacks of deep learning algorithms are partially covered by innovative methods like enlarging the training dataset using data augmentation algorithms and benefiting from the computational power of graphic processing units (GPUs) (LeCun et al. 2015). Table 2-6 shows the studies in automated sewer defect detection using deep neural networks.

Table 2-6. Studies in automated sewer defect detection using deep neural networks

Author(s)	Year	Data Acquisition	Pipe Type	Feature Extractor	Classifier/ Detector	Defect(s)	Assessment Protocol
Kumar et al.	2020	CCTV	VCP	CNN	Faster R-CNN	Deposit, Root	-
Yin et al.	2020	CCTV	VCP	CNN	YOLOv3	Break, Hole, Deposits, Crack, Fracture, Root	-
Kumar et al.	2020	CCTV	VCP	CNN	YOLOv3	Fracture, Root, Deposit	PACP
Hassan et al.	2019	CCTV	Concrete	-	CNN	Crack, Joint, Debris, Lateral	-
Li et al.	2019	CCTV	Concrete, PDE, PVC	-	ResNet18	Deposit, Joint, Break, Deformation	NASSCO
Meijer et al.	2019	CCTV	-	-	CNN	All in Standard Code	European standard coding norm EN 13508-2
Xie et al.	2019	CCTV	PVC	-	CNN	Break, Hole, Deposits, Crack, Fracture, Root	-
Wang & Cheng	2019	CCTV	VCP	-	FCN	Crack, Deposits, Root	-

Author(s)	Year	Data Acquisition	Pipe Type	Feature Extractor	Classifier/ Detector	Defect(s)	Assessment Protocol
Wang & Cheng	2019	CCTV	VCP	-	RNN	Crack, Deposits, Root	-
Moradi et al.	2019	CCTV	VCP	CNN	SSD	Crack, Infiltration, Deposit	PACP
Kumar et al.	2018	CCTV	VCP	-	CNN	Root, Deposit, Crack	-
Cheng & Wang	2018	CCTV	-	CNN	Faster R-CNN	Root, Deposit, Crack, Infiltration	-
Chen et al.	2018	CCTV	VCP		CNN	ROI	-
Moradi et al.	2018	CCTV	VCP		CNN	Crack	-

2.10 Discussion and Gap analysis

Existing challenges in underground infrastructure assessment for municipalities and asset managers motivate the engineering companies to develop new inspection technologies. Considering the introduced achievements in inspection tools, now the inspectors can obtain the data of all types of sewer defects. Moreover, the acquired inspection data has been improved both in quality and quantity. Thousands of hours of CCTV inspection videos are archived and documented in the municipalities and inspection contractors' libraries.

Meanwhile, improvements in machine learning and computer vision techniques have made it easier to propose models in sewer inspection automation and validate them by the available data. Research in sewer inspection automation provides substantial contributions to the detection of a wide range of defects. In this research, the extents of automated sewer defects assessment in the proposed models have been studied in four aspects: detection, localization, evaluation, and algorithm generalization.

Considering the advances in computer vision and especially deep learning algorithms, almost all the detectable defects by CCTV can be detected automatically. In addition, the location of the defects in pipe's cross section has been identified by algorithms like deep object detection frameworks. However, for defects like deposit, complementary tools such as laser scanner employed with CCTV since deposits may be hidden under the waterline. Defects severity evaluation is still the area requiring more study. Some research proposed algorithms based on pixels intensities or saliency models to quantifying the extents of the defected areas. However, their success is still far from an automated model in severity evaluation. The generalization capability of the proposed models is another area for development. The introduced algorithms and frameworks need to be generalized for automated defect detection in various environments and pipe materials. This capability can be improved by generating more comprehensive datasets and train more accurate machine learning models. Furthermore, the frameworks can be designed to be adjustable to various sewer assessment protocols such as WRc, PACP, etc. (Table 2-7).

Table 2-7. Automation of sewer defects assessment (adapted from (Moradi et al. 2019a))

Defect type	Detection	Localization	Evaluation	Algorithm Generalization
Cracks	◙	◙	◘	◙
Joint	◙	◙	○	◘
Deformation	◘	◘	◘	○
Holes	◙	◙	◘	◘
Root	◙	◙	○	◙
Infiltration	◙	◙	○	○
Deposit	◙	◙	○	◙

(◙) achieved, (◘) partially achieved, (○) not achieved

Although there have been remarkable achievements in sewer defect detection automation in recent years, there are still limitations in applicability of these models. Visual data from underground assets like sewer pipes are always prone to poor lighting and varying illumination. Sudden movements and camera pose changes make it harder to keep consistency in the data analysis. To add to this, sewer inspection videos include a wide range of frames without any valuable information like underwater, information, start, and ending frames. Thus, data cleansing and video preprocessing seems to an inevitable part of the frameworks.

In addition, quality of image data from sewer pipelines is always substandard due to various noises and existing occlusions. The images are always supplemented with occluded and intermittent background. Defect features and templates may diverge by an infinite range regarding the camera pose, distance, defect size, pipe wall cleanness, etc. Thereby, the detection algorithms require larger datasets for training to improve their generalization capability. Furthermore, the whole assessment process is still operator interactive since measure and score the defects severity yet are dependent on operator's judgment.

The recent deep learning based frameworks offered a leading advantage in extracting image features automatically. Feature selection and extraction are the most important steps in a classifier design since features need to provide enough discriminatory information, and at the same time, be easy to compute. In traditional machine learning algorithms, the defect features have been selected based on geometric form and size of defects existing in sewer image. However, infinite patterns

and templates of defects in sewer images made it challenging to propose a general feature for each type of defect.

Additional attribute of image analysis techniques is the usage of color channels and their intensities. Generally, color images are converted to one channel grayscale to decrease the computational cost and accelerate the calculations. However, pixels color intensities may be useful in differentiating among the defect and image background. Also, pipe wall color can confuse the classifier in detection of some types of defects like infiltration.

The proposed deep learning algorithms for sewer defect detection presented a promising performance in detection accuracy but these algorithms are expensive to train and exploit. Large training labeled datasets and powerful computational hardware are main downsides of deep learning algorithms. The detection speed is another concern since employing a more accurate and deeper framework reduces the real-time detection capacity and speed. Small objects in pipe like holes is another difficulty for the existing defect detection models. A tradeoff between speed and accuracy may result in missing some of the defects which are too small to be detected in limited processing time.

In conclusion, application of computer vision in sewer inspection automation is highly dependent on the provided visual data by CCTV. The algorithms analyze what has been recorded by CCTV cameras and whatever cannot be acquired by the camera cannot be evaluated by the computer vision algorithm. The quality of the performed video data needs to be standardized and camera angle and pose in the pipe should be justified precisely based on the procedures is available assessment protocols. The pipes required to be lighted properly to minimize the illumination fluctuation in the images. Considering all the mentioned suggestions, and future developments in algorithms and hardware facilities it is expected sewer pipelines assessments would be automated and result in a substantial decrease in inspection time and cost of sewer pipes.

Chapter 3 : Methodology and Model Development

In this chapter, the proposed research methodology is presented where the developed models in each step are highlighted. The proposed methodology aims to support the sewer inspection process by automating the CCTV video analysis and defect detection. The whole framework is divided into two main phases: identifying the defective frames, and defect detection and classification. In this chapter, the proposed methodology has been explained thoroughly. The development of models, including anomaly detection, defect detection, and classification algorithms, are presented. In the following sections, first, collecting the different types of data used in model development is described. In the next parts, each step of the proposed methodology is demonstrated.

3.1 Proposed methodology

The current practice in sewer inspection is first to detect the areas which are suspected to be a defect. This step is usually conducted by the operator and is highly dependent on his or her vision and ability of recognition. Therefore, that is subject to the operator's skillfulness, fatigue, illumination, camera pose, water level in the pipe, and other operational conditions. Based on a study conducted by Dirksen et al. (2013), in manual sewer inspection 25% of defects are missed by the operator. So, it is important to minimize human error to be able to detect all potentially defected areas in inspection video frames. In this research, at the first stage, sewer CCTV videos would be analyzed to recognize suspicious frames, which may include any types of defects (anomalies).

In current manual practice, after identifying suspicious regions, the collected evidence would be analyzed by the operator on-site or later by a certified expert in the office to detect the defects and classify them based on type, criticality, and location. This phase is also reliant on the operator and is subjective. Thereby, the second part of the proposed model aims to analyze the extracted anomalous frames using deep learning techniques to classify and detect the defects based on their type. The training and detection process is illustrated in section 3.3. Figure 3-1 shows the proposed methodology.

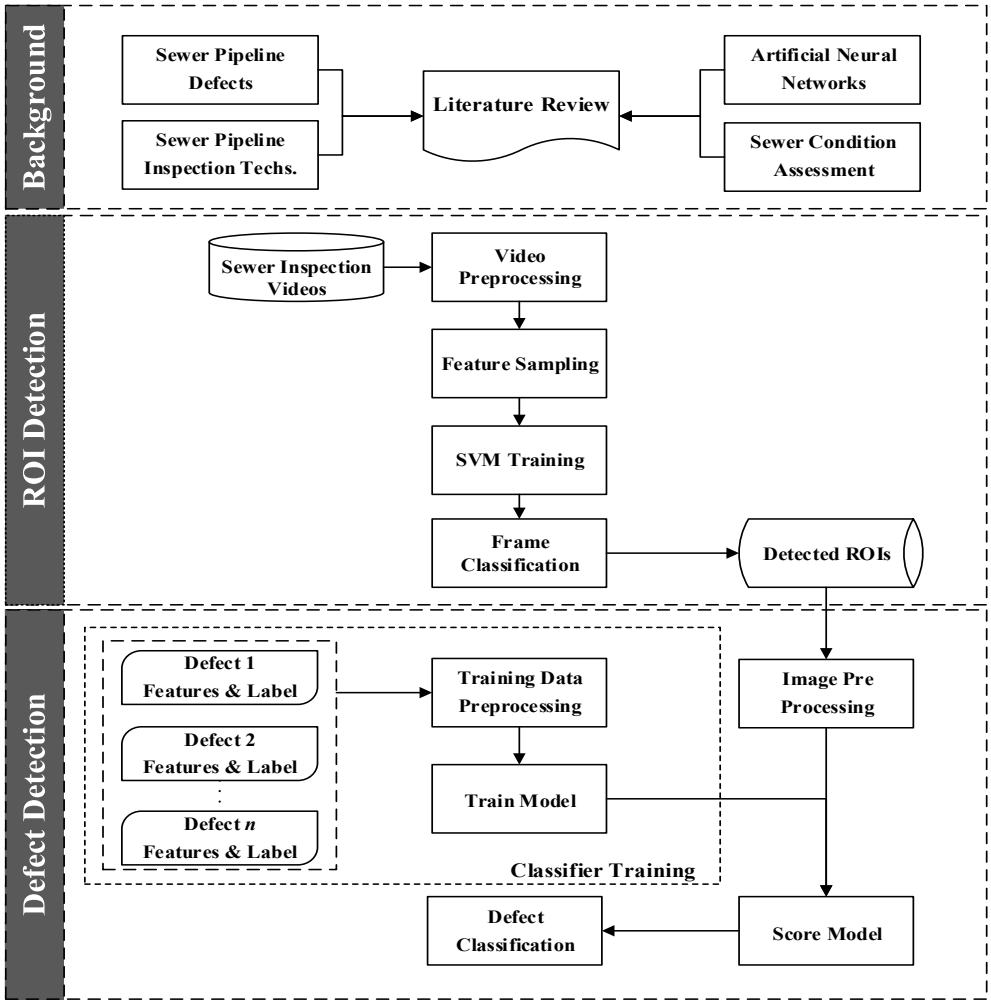


Figure 3-1. Overall view of proposed methodology

3.2 Data collection

In the data collection stage, required datasets for the development of the models are provided. These datasets comprise data extracted from CCTV inspection videos and reports for two existing sewage networks in Qatar and Canada, and existing literature, as shown in Figure 3-2.

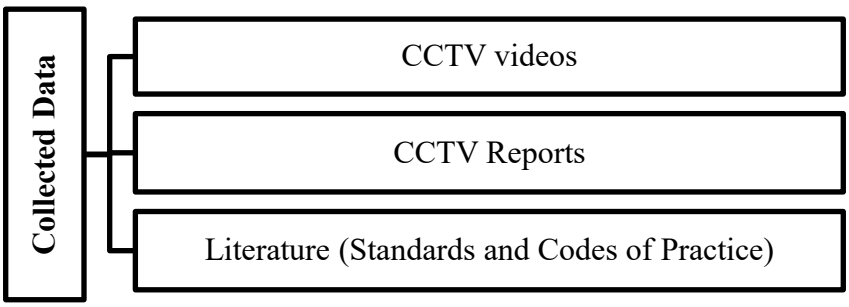


Figure 3-2. Collected data types used in framework development

3.2.1 Input datasets

Each dataset is used in a specific stage of model development. Acquired CCTV footages from Laval and Qatar are used in the development of the first step as extracting various features of video frames and in training and testing the classifier in the following. Hours of inspection CCTV videos have been analyzed and modified to fit properly as algorithms input. These videos are used in classifier training and testing in the first step. Also, the frame localization step is provided by analyzing text information in video frames. The anomalous frames in the inspection videos and defect images extracted from CCTV inspection reports feed as input to neural networks in both the training and testing stages of the model. Features are extracted from available images, and image processing techniques are developed based on the CCTV inspection videos specifications (illumination, quality, etc.).

Also, the videos and reports were analyzed to extract the defects and generating the relative datasets. The videos from City of Laval were investigated to capture the sewer defects among the video sequences and creating image datasets for four types of defects as cracks, infiltration, joint displacement, and deposits. These datasets were enlarged by adding the categorized defect images in the reports from Qatar.

All the images were reviewed and studied to ensure correct classification and then the prepared dataset was labeled for different needs. The CCTV footage was categorized into normal videos (containing no defect) and videos with anomalies. A dataset of images was labeled for the text detection model. Image labels include English numbers and alphabets. Moreover, the images of defects were labeled both for the type of defect and relative location of the defect in that image. Details of image labeling are illustrated completely in the following sections.

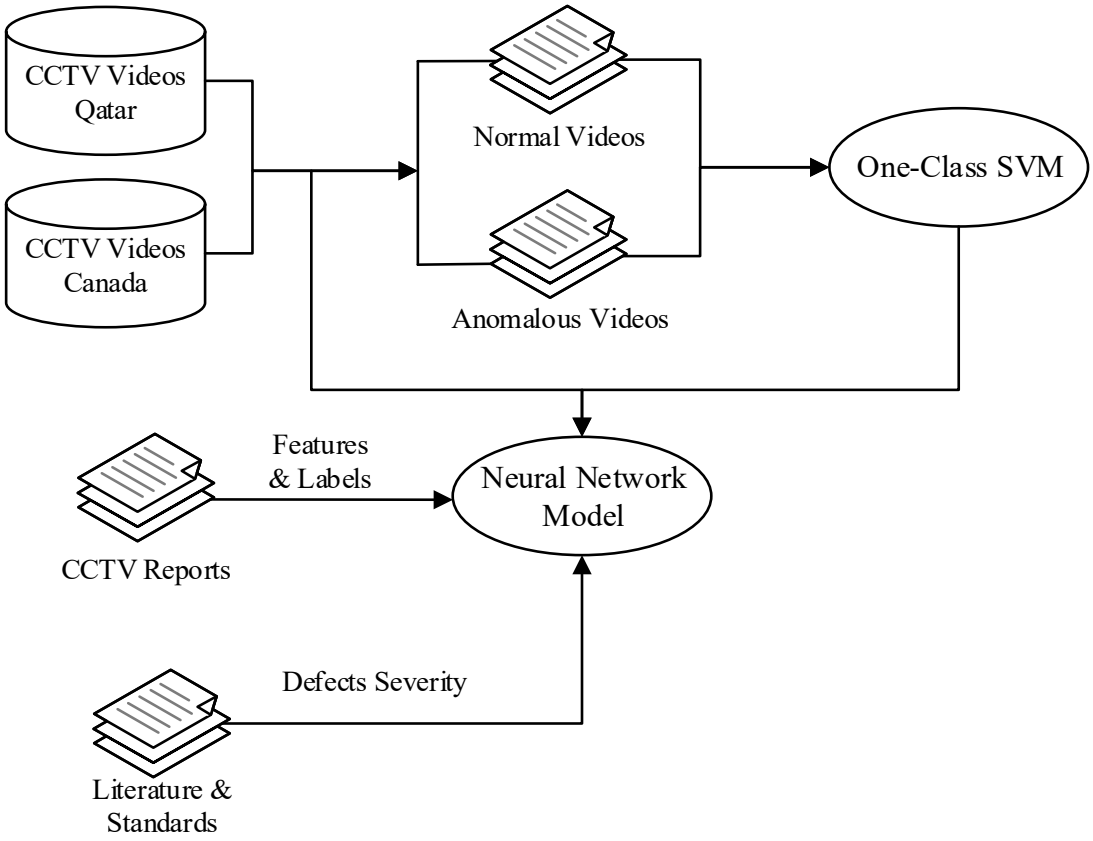


Figure 3-3. Interaction between Different Datasets in Building Models

3.2.2 Data augmentation

Deep learning algorithms are thriving for data, and the more data provided for training, the better performance can be achieved. Training the network with small datasets may result in overfitting and lack of generalization capability for the developed models. So, data augmentation tends to generate new training data by artificially increasing the size of the dataset. A common and simple practice in image data augmentation is classical image transformations such as shifting, resizing, rotating, and cropping. The image transformations can be fused with image color adjustments such as histogram equalization, contrast enhancement, brightness enhancement, sharpening, and blurring.

In this research, the provided dataset from sewer defect images has been augmented by various image transformation techniques such as resizing, rotation, flip, shifted in different color channels shifts, and added Gaussian noises. As result, the dataset size was increased considerably. Each defect sample set contained 3000 images. In some defects such as cracks, subcategories are ignored so longitudinal, diagonal, and complex cracks are all considered as a crack in the implementation. Figure 3-4 shows different image transformations applied to the dataset.

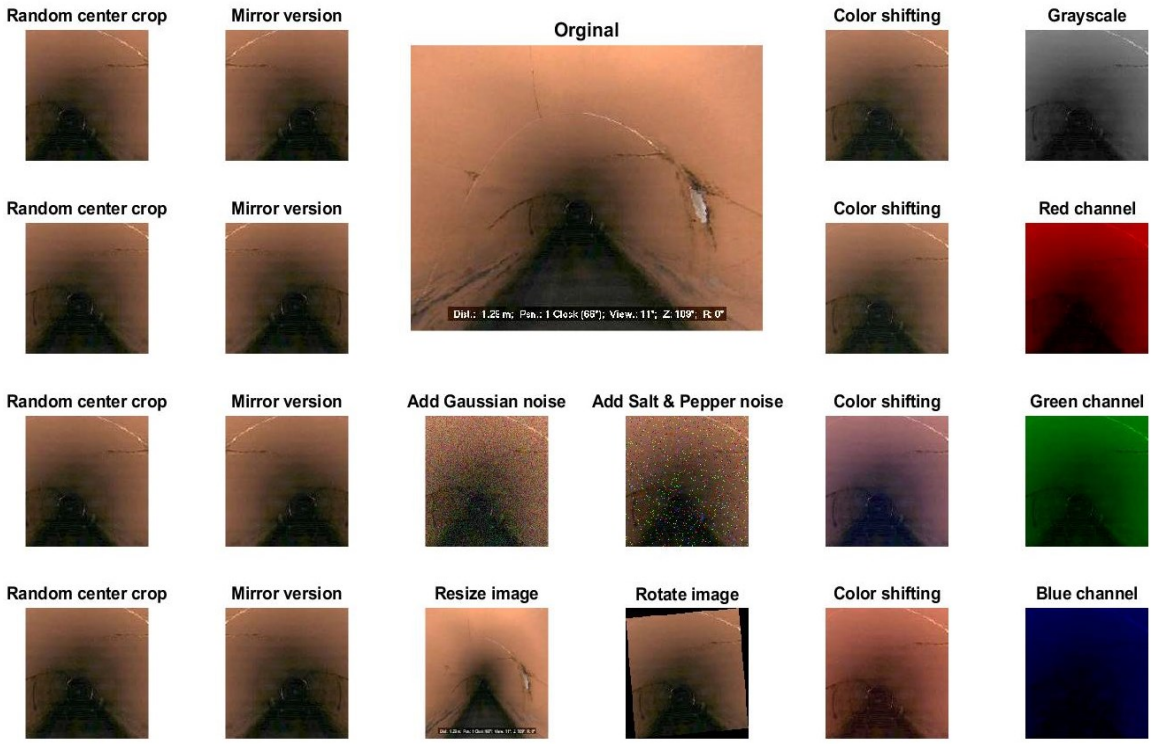


Figure 3-4. Example of various image data augmentation

3.3 Anomaly detection

This chapter is a marginally modified version of "*Automated Anomaly Detection and Localization in Sewer Inspection Videos Using Proportional Data Modeling and Deep Learning–Based Text Recognition*" published in *Journal of Infrastructure Systems* (Moradi et al. 2020) and has been reproduced here.

Due to the exceptional internal conditions of sewer pipes, the recorded videos are not easy to analyze. Visual characteristics of the objects could be varying depending on camera pose, illumination condition, type of sewer, pipe material, and pipe diameter. Waterline fluctuations and sudden movements of camera are also affecting the uniformity of visual inspection. All the mentioned factors in addition to the limitations of CCTV inspection technology lead to unstructured and inconsistent data which can be analyzed by simple computer vision tools. Thereby, to develop a robust model to automate sewer inspection, complex methods should be integrated. In this research, an innovative framework is proposed for anomaly detection and localization in CCTV videos. Figure 3-5 shows the general overview of the proposed framework.

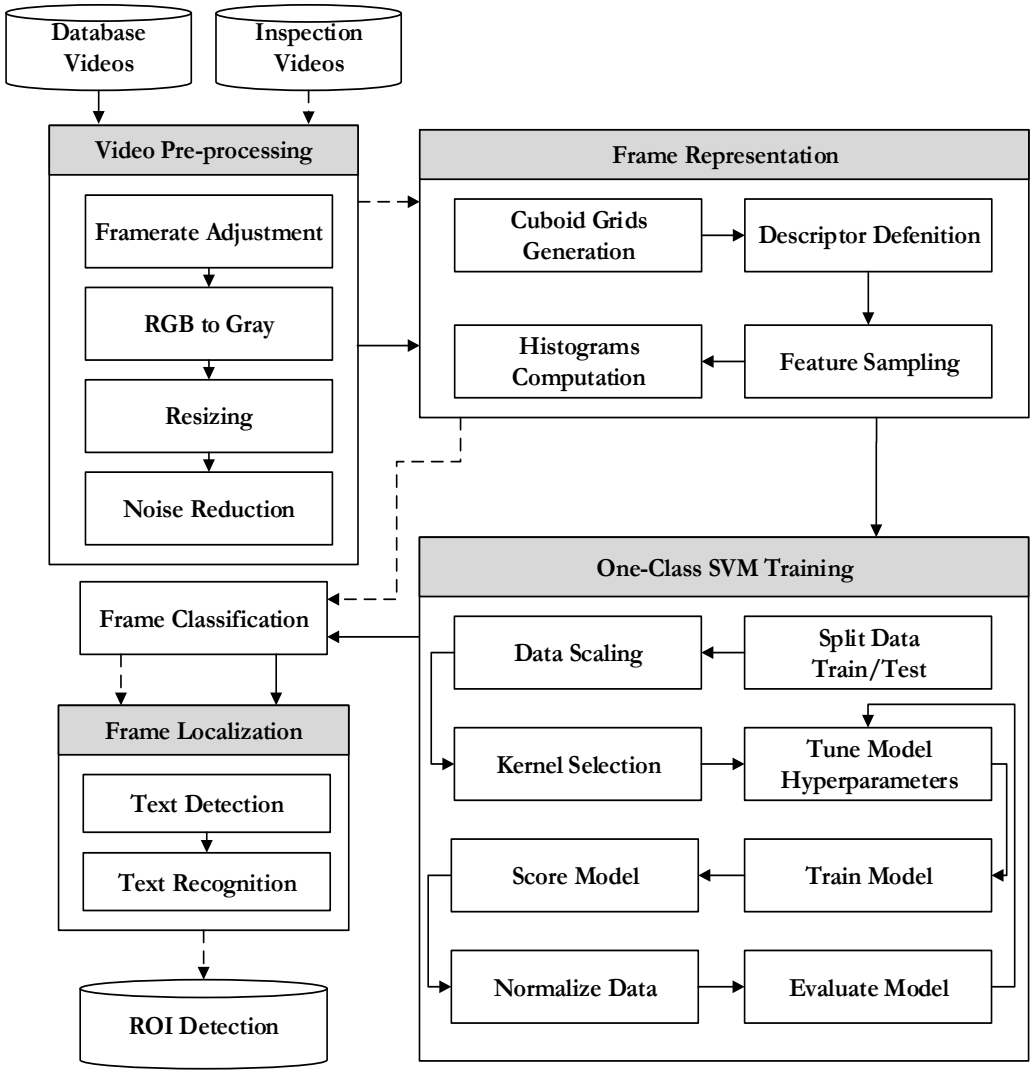


Figure 3-5. ROI detection and localization (adapted from (Moradi et al. 2020))

The input dataset needs to be cleaned and a set of operations are applied to the images to prepare them for model input. Considering the slow speed of the tractor (3-10 m/min) which is conveying the CCTV in the pipe, and the frame rate that inspection videos are recorded (i.e., 30 fps), each second of the recorded footage includes almost the same information and not many image features variations would be observed. So, in the first step, the frame rate is decreased. Thereby, the input dataset will shrink and there would be a remarkable reduction in input data.

In the sewer pipeline, most of the defects can be patterned by binary (light/dark) primitive shapes (Moradi et al. 2019a), and image colors are not that much helpful in detecting the edges and other features of the defects in the sewer pipe. So, to reduce the number of calculations and increase the speed, the three channels RGB inspection video frames were modified to grayscale level of the pixels. Furthermore, the number of input pixels for each image is optimized by rescaling and resizing the input image size. These three operations, frame rate adjustment, greyscale conversion,

and rescaling, intend to increase the computational speed. As mentioned in previous sections, obtained image data from sewer pipes are too noisy, so it is required that the images quality be enhanced by noise reduction operators.

3.3.1 Frame representation

Sewer inspection videos include a large number of frame sequences with normal pipe conditions and just short durations of defected pipe frames. So, the defected (anomalous) frames features need to be captured densely and rich enough to construct proper feature vectors. The proposed model generates a grid of cuboids from both normal (healthy) and anomalous (defected) inspection video frames. The dense features of the image cuboids are captured using an innovative image representation method based on scale invariant feature transforms (SIFT) (Lowe 1999, 2004) and inspired from 3D SIFT approach which is an extension of the SIFT descriptor developed by Scovanner et al. (2007). The image features captured by SIFT descriptor are able to remain unchanged to rotations and scaling transformations, robust to perspective deviations, and illumination variations (Lowe 2004). In this research, a modified 3D SIFT approach utilized to encode the sewer inspection videos information.

Scale-space Extrema Detection

The first step in the approach is to identify image key points. Image pixels are surveyed through all scales and image pixels to locate potential key points using a cascade filtering approach (Lowe 2004). In the first step, a set of image volumes with different sizes are generated. Then, a 3D Laplacian of Gaussian (LoG) filter with different σ values is applied to the generated image volumes. A lower σ Gaussian filter allocates to the small angles with high values, whereas a Gaussian filter with larger σ corresponds to a larger corner. In this research, image volumes were scaled to three octaves with three different σ levels and $\sigma_0 = \sqrt{2}$ as initial σ based on Lowe (2004). LoG requires a large number of computations, so a modified function of the Difference of Gaussians (DoG) is applied on different image volume scales in the 3D SIFT algorithm, which is a close estimate of the LoG (Lowe 2004). Finally, the determined local maxima across all scales and spaces are a group of (x,y,z,σ) values that signify the potential key points at (x,y,z) and σ scale (Figure 3-6).

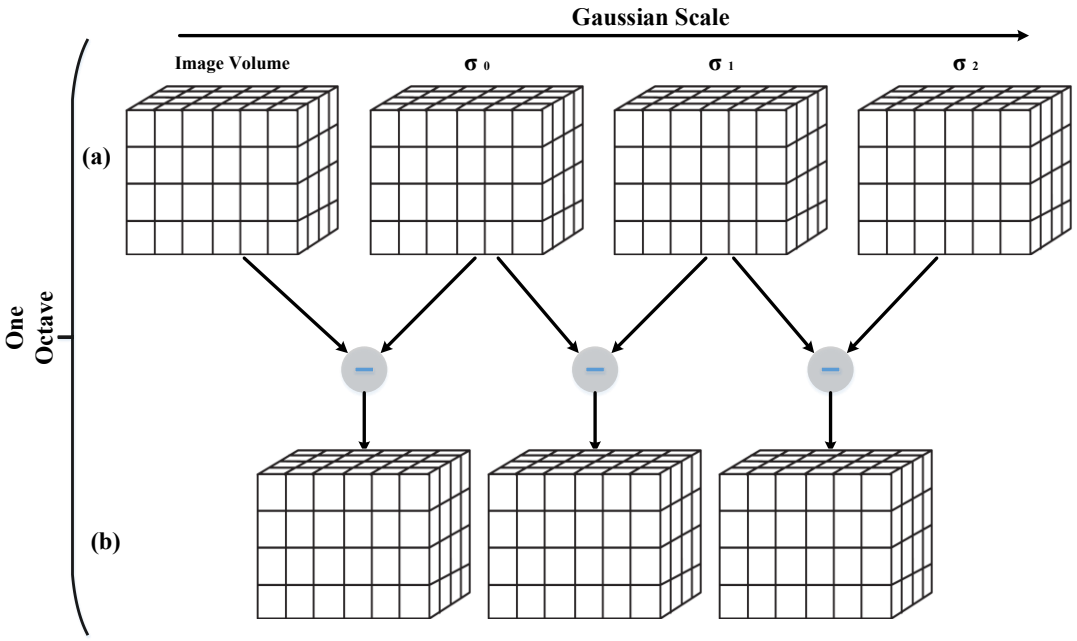


Figure 3-6. (*a*) A set of scale space images repeatedly convolved with Gaussians, (*b*) Subtraction of adjacent Gaussian images to produce the difference-of-Gaussian (DoG) images (adapted from (Moradi et al. 2020)).

Then, "each represented potential keypoint was compared to its 26 neighbor pixels in the current volume and two other sets of 27 pixels of the upper and lower space volumes. If the pixel was a maximum or minimum among the other pixels, it was assigned a value of 1; otherwise, it was 0" (Moradi et al. 2020). Figure 3-6 presents the maxima and minima identification among the generated image volumes.

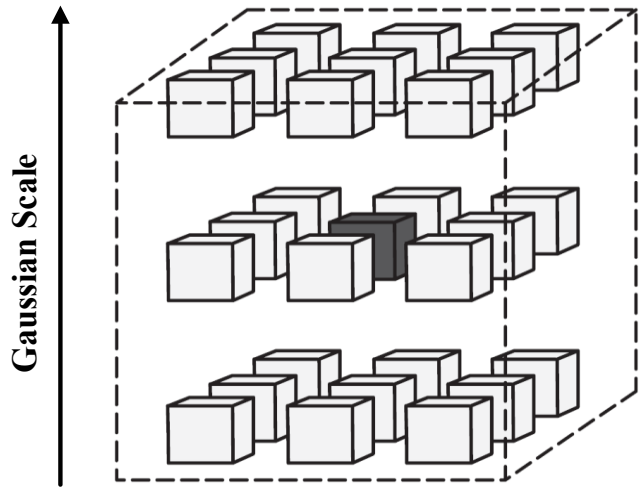


Figure 3-7. Maxima and minima of the difference-of-Gaussian image volumes by comparing a pixel to its neighbors (adapted from (Ni et al. 2009))

Keypoint Localization

The identified maxima and minima points are represented as the potential key points. However, the potential key points which do not include important information are filtered. More accurate extrema locations are extracted using Taylor series, considering a threshold value of 0.03 (Lowe 2004), and low contrast key points that their intensities are lower than the threshold would be excluded. DoG function has a higher effect on edge pixels, and they are uninvolved by a 3×3 Hessian matrix (H) to calculate the principal curvature (Equation 3-1).

$$H = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{yx} & D_{yy} & D_{yz} \\ D_{zx} & D_{zy} & D_{zz} \end{bmatrix} \qquad \text{Equation 3-1}$$

An eigenvalue with the largest magnitude to the smallest one was considered as H ($\lambda_1 < \lambda_2 < \lambda_3$) and the sum of the eigenvalues from the trace of H and their product from the determinant computed as follows (Ni et al. 2009) (Equation 3-2 & 3-3):

$$\text{Tr}(H) = D_{xx} + D_{yy} + D_{zz} = \lambda_1 + \lambda_2 + \lambda_3 \quad \text{Equation 3-2}$$

$$\text{Det}(H) = D_{xx}D_{yy}D_{zz} + 2D_{xx}D_{yy}D_{zz} - D_{xx}D_{yz}^2 - D_{yy}D_{xz}^2 - D_{zz}D_{xy}^2 = \lambda_1\lambda_2\lambda_3 \qquad \text{Equation 3-3}$$

Considering *r* as the ratio between the largest magnitude eigenvalue and the smaller one, if the $r = \frac{\lambda_3}{\lambda_1} < r_{max}$, then the feature point is acceptable. If $\alpha = \frac{\lambda_2}{\lambda_1}$, then

$$\frac{\text{Tr (H)}^3}{\text{Det (H)}} = \frac{(r+\alpha+1)^3}{r\alpha} \qquad \text{Equation 3-4}$$

Lowe (2004) used 10 for the value of r, which excludes key points with a ratio less than 12.1. eventually, the low-contrast keypoints and edge keypoints are filtered and the accurate key points are determined as:

$$\frac{\text{Tr (H)}^3}{\text{Det (H)}} < \frac{(2r_{max}+1)^3}{r_{max}^2} \qquad \text{Equation 3-5}$$

Orientation Assignment

Key points orientations are calculated based on local image gradient direction in each key point and to accomplish image rotation invariance, the overall orientation of each neighborhood is determined. For the relative scale, the pixels neighborhood around the keypoint location is taken to determine the gradient magnitude and two directions of each pixel (space-time) in the neighborhood. Polar coordinates of the key points were calculated by one magnitude and two angles using the equations introduced in (Scovanner et al. 2007) :

$$M_{3D} = \sqrt{G_X^2 + G_y^2 + G_t^2} \qquad \text{Equation 3-6}$$

$$\Phi = \tan^{-1}(G_t / \sqrt{G_X^2 + G_y^2}) \qquad \text{Equation 3-7}$$

$$\theta = \tan^{-1}(G_y/G_x) \qquad \text{Equation 3-8}$$

Where G_x, G_y, and G_t are gradients in Cartesian coordinates. Two distinct orientation histograms will be quantized using θ and φ and be weighted by the magnitude M_{3D} which is always positive. Also, φ would be in the range $(-\frac{\pi}{2}, \frac{\pi}{2})$ and θ would be in the range $(-\pi, \pi)$. Thereby, the direction of the gradients of keypoints is presented by two values (φ and θ (Scovanner et al. 2007).

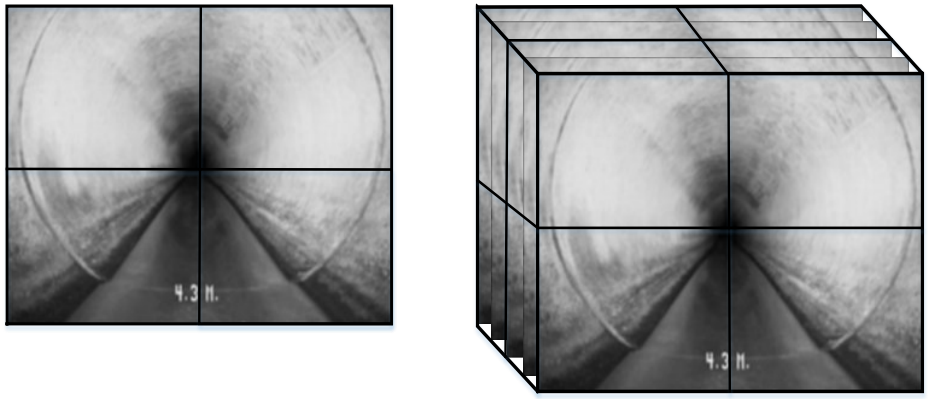


Figure 3-8.Example of a cuboid with 3D SIFT descriptor in sub-regions (adapted from (Moradi et al. 2020))

"For each 4×4×4 sub-region of a cuboid, three-dimensional gradient-based features were computed based on pixel luminance values. Consequently, each sub-region was quantized as a 12-bin histogram, including the values of φ in four bins and the values of θ into eight bins. For each 3D sub-region, the orientation was accumulated into an 8x4 histogram and configured as 4×4×4 sub-histograms" (Moradi et al. 2020).

3.3.2 Anomaly detection using Support Vector Machine

SVM is a powerful statistical machine learning model that is able to perform linear or nonlinear regression, classification, and anomaly detection. SVMs determine the best possible splitting line, plane, or hyperplane among two classes or more. For each class in the dataset, SVM maximizes the closest data points distance and the data points lying on the boundaries are called support vectors (Cortes and Vapnik 1995). The maximization of data point margins among the different classes datapoints leads to the best separating boundaries and improve the model generalization. The model considers a weight for the data points which are sat on the wrong side and to minimize their effect on the classification performance the weight tends to be lessened.

SVMs offer some advantages over other supervised machine learning algorithms when dealing with small to medium size training datasets. In comparison to high parameter enabled models like neural networks, SVMs are faster both in training and classification phases. The main benefit of SVMs is their ability to cover high dimensional datasets efficiently, and classification performance does not affected by the size of feature space (Joachims 1998).

In sewer pipes most of the frames do not include any type of defect (normal frames), while short temporal durations of the frame may contain defects (anomalous frames). In this research, the sewer inspection video frames are classified into two classes of normal and anomalous frames. The features of anomalous frames are not known, and anomalies can be appeared in any shapes. On the other hand, the normal frames features are known, and their features can be extracted to generate the relative training dataset. Thus, normal and anomalous frames can be classified using a one class classifier which is trained by normal frames features. So, any datapoints of the input data would be classified into normal or anomaly based on its comparative position to the dataset used in the training. In this research, one class SVM is introduced to perform anomaly detection among sewer inspection video frames.

One class SVM is a common technique in anomaly detection proved to be robust tools in anomaly detection in high dimensional and large scale CCTV frames (Erfani et al. 2016; Moradi et al. 2020; Yang et al. 2019). In the training of OC-SVM, the relative distribution of normal data is modeled and the data is differentiated by a specific kernel function that plots the input space to a higher dimensional feature space (Erfani et al. 2016). In the developed framework, the provided dataset from the sewer frames sequences, is split into training and test datasets. Large values of the attributes are required to be scaled to prevent missing smaller numeric ranges. Moreover, large numeric ranges result in calculation difficulty.

In this research, OC-SVM algorithm inspired by Schölkopf et al. (2001). The proposed algorithm maximizes the decision boundary margin among the normal data points and the origin and precludes expensive calculations for high-dimensional datasets (Schölkopf et al. 2001). To establish the support vector, the cost function is (Schölkopf et al. 2001):

$$\text{Minimum } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i - \rho \qquad w \in F, \xi \in R, \rho \in R \qquad \text{Equation 3-9}$$

Where $C = \frac{1}{N\vartheta}$ and ;

$$subject\ to \quad w.\varphi(x_i) \geq \rho - \xi_i, \quad \xi_i \geq 0.$$

“Where N is the number of the data points in normal dataset, ν is a regularization parameter, and ξ_i is the slack variable for point $\mathbf{x_i}$ that allows for some anomalies to locate outside of the decision boundary and $\xi=[\xi_1,\ldots,\xi_N]$. The parameter *ν* rules the fraction of normal data that possibly will be classified as outliers, whereas *w* and *ρ* are the parameters which determine the decision boundary and are the target optimization problem variables. *x* and φ the original data sets into feature space which is a higher dimensional space to linearly separate the non-separable data sets and $K(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j)$ is the kernel function” (Moradi et al. 2020).

To limit the variable dimensions, the cost function can be translated into a dual problem as follows (Schölkopf et al. 2001):

$$\min \quad \sum_{i\,j}^{n} \alpha_i \alpha_j k(x_i, x_j) \qquad \alpha \in R \qquad \text{Equation 3-10}$$

$$subject\ to: \sum_{i=1}^{n} \alpha_i = 1$$

$$0 \leq \alpha_i \leq \frac{1}{vn}$$

as α denotes the Lagrange multiplier and $k(x_i, x_j)$ represents the dot product of x_i and x_j vectors. To more demonstration of the algorithm can be found in Schölkopf et al. (2001)., The decision boundary is determined using Lagrange techniques and utilizing a kernel function (Shahid et al. 2015):

$$f(x) = sgn((w \cdot \phi(x_i)) - \rho) = sgn(\sum_{i=1}^{n} \alpha_i K(x, x_i) - \rho) \qquad \text{Equation 3-11}$$

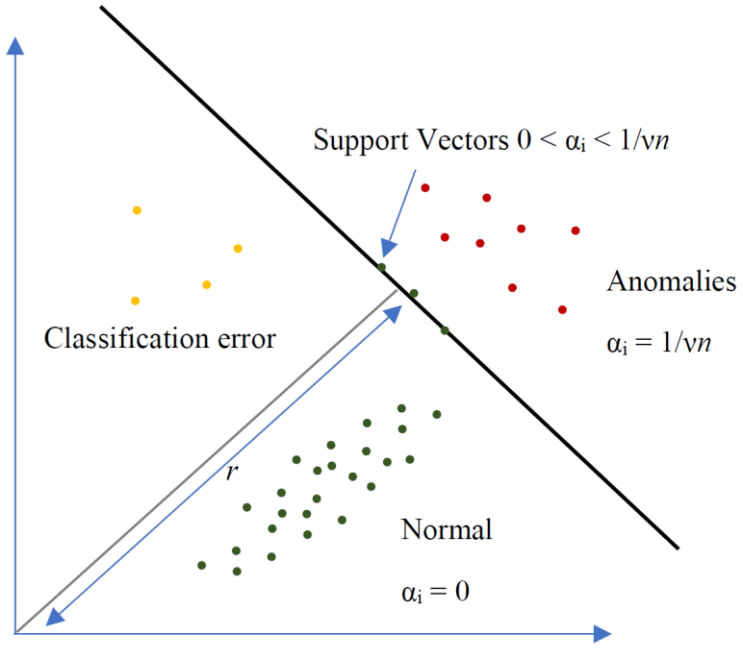


Figure 3-9. General scheme of One-Class SVM (adapted from (Shahid et al. 2015))

One of the most significant aspects of the SVMs is the kernel technique. Kernel parameters impact the classifier's performance and its generalization capability. The commonly employed kernel functions are as follows (Yin et al. 2014):

Linear kernel: $K(x_i, x_j) = x_i^T x_j$

Polynomial kernel: $K(x_i, x_j) = (\gamma x_i^T x_j + c)^p$

Radial basis function kernel: $K(x_i, x_j) = \exp(-\gamma \|x - \acute{x}\|^2) . \gamma$

Sigmoidal kernel: $K(x_i, x_j) = tanh(k x_i^T x_j - c)$

where γ and c are constants, σ is the width of the Radial Basis Function (RBF) kernel, and p is the degree of the polynomial.

3.4 ROI localization

Generally, sewer pipeline inspection is conducted in segments of the pipeline. A segment is from starting manhole till the end manhole and can be between 5 to 90 meters. The CCTV camera is mounted on a tractor and the covered distance from the starting manhole is measured and among some other information are indicated as text subtitles in the foreground of the recorded footages. These text subtitles can instruct operational information about the inspection such as inspected pipe address, GIS codes, pipe material, operator's name, and the distance from the starting manhole.

In this research, the distance information is used to locate the identified anomaly or regions of interest (ROIs) in the pipe segment. An innovative end-to-end text detection and recognition framework is proposed for ROI localization. However, applying text recognition algorithms like optical character recognition (OCR) on the recorded image frames from internal sewer pipeline is a quite difficult task. Sewer images are too noisy and include occluded background and varying illumination. So, the proposed approach first detects the text in image and then recognizes the text characters. Figure 3-10 illustrates an outline of the proposed framework.

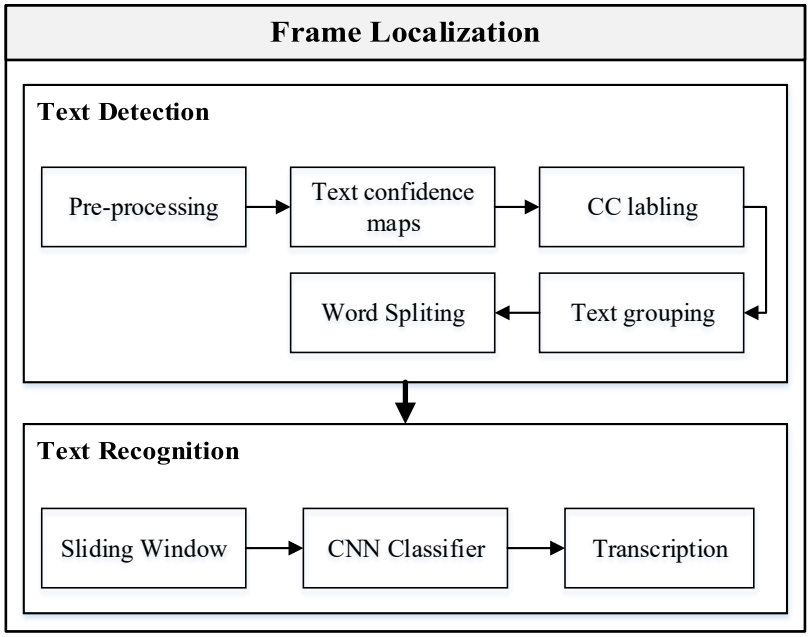


Figure 3-10. Frame localization framework (adapted from (Moradi et al. 2020))

To recognize a text in an image, first its location is required to be detected. There are two main approaches in text detection: region-based and texture based (Epshtein et al. 2010). In region-based approach, a neighborhood of image pixels are grouped as Connected Components (CC) that distinguish character candidates and exclude non-text regions by geometric constraints (Epshtein et al. 2010; Opitz et al. 2014). Alternatively, in texture-based approaches, the image text texture is defined as a distinguishable feature, and to detect the texts, the extracted text features are classified by a classifier (Lee et al. 2011; Pan et al. 2009). Similarly, text recognition algorithms utilize a

classifier to classify segmented CCs or image features such as Histogram of Oriented Gradients (HOG) to recognize the text characters (Opitz et al. 2014).

In dealing with cluttered images, region based approach is proved to outperform texture based approach (Chen et al. 2011; Epshtein et al. 2010; Opitz et al. 2014). The input images from sewer inspection videos are too noisy and the background is highly patterned. Thereby, the text texture edges cannot be distinguished plainly from the background or the neighbor pixels. In this research, region-based approach is used to extract the relative CCs of texts in the sewer images and the steps of the approach are explained thoroughly in the following sections.

3.4.1 Text detection

A sequence of pre-processing operations is applied to the input images to enhance the contrast, sharpen the image, and to adjust pixel intensity values. Then, text confidence maps are generated using Maximally Stable Extremal Regions (MSER) (Matas et al. 2004). In the next step, identified CCs are labeled, and non-text regions are filtered. Finally, text candidates are assembled, and words in detected text lines are identified.

MSER regions extraction

MSER a robust region detector algorithm particularly when it comes to perspective change, scale, and brightness variations. Technically, in an image, pixels of text regions offer consistent color and intensity, and have a substantial difference from the background. MSER algorithm can identify text regions accurately. However, MSER is sensitive to image blur and in low-resolution images small letters are hard to detect (Chen et al. 2011). In this study, MSER algorithm is altered using an edge-enhancement operator to overcome the mentioned drawbacks.

Sewer inspection images are likely to have constructional noises like white noises due to the inner environments of pipe. So, the edges of extremal regions are enriched by an improved Sobel edge detector proposed by Wenshuo Gao et al. (2010). Sobel detector determines the edges in the image and feeds them to MSER algorithm and pixels found outside of the detected edges by Sobel would be excluded. Figure 3-11 illustrates the edge enhanced text confidence maps resulted from MSER.

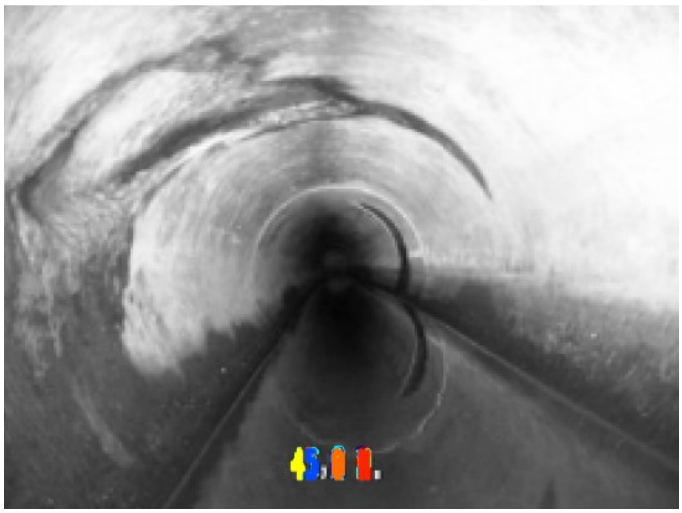


Figure 3-11. Edge-enhanced MSER detection (image adapted from (Moradi et al. 2020))

Connected component labeling

The output of edge enhanced MSER is a set of CCs from image foreground, which are considered as text candidates and a rule-based approach is employed to filtered out the non-text regions among them. To label CCs as text and non-text geometrical characteristics of the detected regions are performed to label CCs (Equations 3-12 to 3-14) (González et al. 2012; Li and Lu 2012):

$$Aspect\ ratio = \frac{\max(width,height)}{\min(width,height)} \qquad \text{Equation 3-12}$$

$$Compactness = \frac{area}{priemeter*priemeter} \qquad \text{Equation 3-13}$$

$$Solidity = \frac{area}{convex\ area} \qquad \text{Equation 3-14}$$

By the proposed ruled based approach, in the first step all the identified CCs containing numerous holes and very large or very small aspect ratio are removed. Then a threshold for compactness and solidity of regions pixels ratios is defined, and detected regions with compactness and solidity ratios lower than the threshold are simply identified as non-text regions and will be discarded.

Remove non-text regions using Stroke Width

After labeling the identified CCs, still there are some non-text regions remained that need to be filtered. To identify and eliminate the remained non-text regions stroke width transformation as another metric is utilized (Chen et al. 2011; Epshtein et al. 2010; Li and Lu 2012). "" (Moradi et al. 2020). Stroke width size, max stroke width, and stroke width variance ratios are calculated using the following calculations (Equation 3-15 to 3-17):

$$Stroke\ width\ size\ ratio = \frac{Stroke\ width}{\max(height,width)} \qquad \text{Equation 3-15}$$

$$Max\ stroke\ width\ ratio = \frac{Max\ stroke\ width}{\max(height,width)} \qquad \text{Equation 3-16}$$

$$Stroke\ width\ variance\ ratio = \frac{Stroke\ width\ variance}{Stroke\ width} \qquad \text{Equation 3-17}$$

As shown in figure 3-12, in a text region in image the variations in stroke width of lines and curves are limited over most of the region.

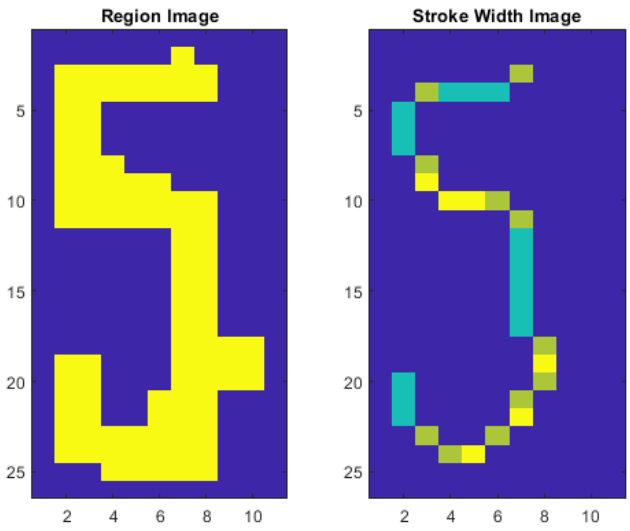


Figure 3-12. Stroke width of text regions

Text line formation and merge text regions

Another technique for text region detection is text line formation. Generally, text regions in an image form a line as they come into view one after each other, so text line is a key indicator of text existence. Text line detection tends to eliminate false positives identified in the previous steps. The text region candidates are being evaluated in a pairwise comparison to merge the CCs into text lines. Theoretically, word characters come within a single text line and share similar attributes such as stroke width, letter height, intensity, and size (Chen et al. 2011). The text information in sewer images appears in the form of straight lines. However, in sewer images, repeated patterns cause a high-level of false positives. The text lines with high level of repetitive objects would be identified as false positives and would be rejected. In the proposed approach a template matching algorithm introduced by Chen et al. (Chen et al. 2011) used to filter out the false positives amongst the recognized character candidates.

Ultimately, the final text detection outcomes will be highlighted as unified text regions (Figure 3-13).

Figure 3-13. Highlighted detected texts in a sewer image (adapted from (Moradi et al. 2020))

3.4.2 Text recognition

In the next step of the proposed frame localization framework, the detected text regions are fed into a text recognition algorithm to define the text characters. The main three steps in text recognition methods are pre-processing, character segmentation, and character recognition (Long et al. 2018). As mentioned in earlier sections, sewer inspection images have complex background and patterned environments and segmenting the text characters is very challenging. In the latest researches, character segmentation has been avoided using methods such as Connectionist Temporal Classification (CTC) (Graves et al. 2008; Yin et al. 2017) and Attention Mechanism (Long et al. 2018). In this research, the CTC recognition method introduced by Yin et al. (2017), is utilized since it is more straightforward to employ.

In CTC approach, the characters feature maps in the detected text regions are extracted by convolutional 1D sliding window. “The sliding window's height is justified based on detected text box height, and its width is based on the width of characters in the text image to ensure that the characters are covered thoroughly. In the next step, the extracted features are fed to the classifier, which is trained to predict the label for input features. In the end, the predicted characters are decoded in the transcription step (CTC layer) to recognize the word” (Moradi et al. 2020).

In this research, text character classification is performed by a 5-layer architecture Convolutional Neural Network (CNN) (figure 9). The highlighted text regions are masked out input to the CNN and resized to 32 × 32 pixels. The inputs pass through convolutional layers with a 3×3 receptive field, convolution stride one, and rectified linear unit (ReLU) as the activation function. A max-pooling layer with 2×2 window and stride 2comes right after each convolutional layer to downsize the number of parameters in the network. At the top of the network the feature vectors are flattened by two fully connected layers and the probability of each character class is determined by a softmax layer.

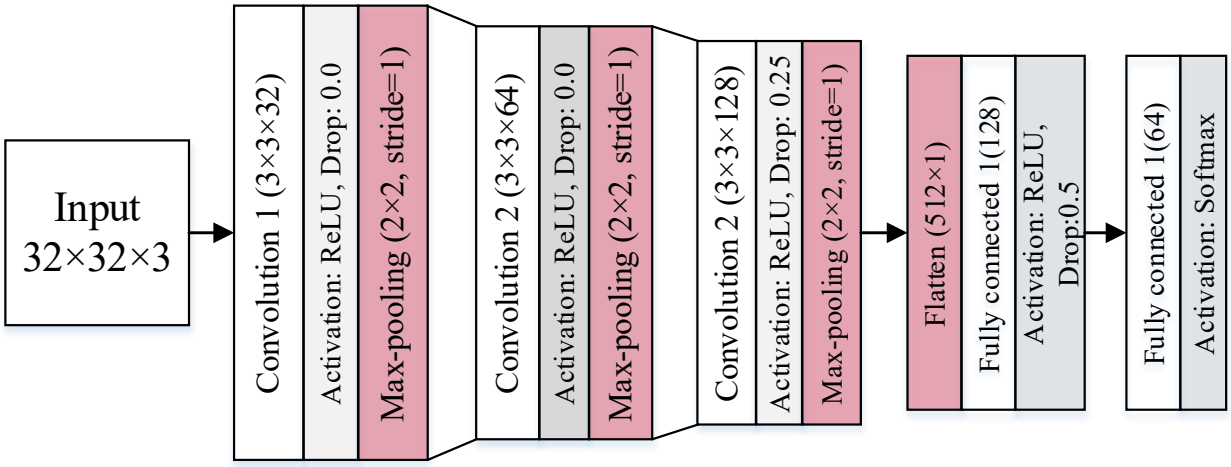


Figure 3-14. Character classification model architecture (adapted from (Moradi et al. 2020))

For each text window, the recognized characters by the classifier are transcribed into a sequence character label using CTC method. CTC method omits training data pre-segmentation by maximizing the conditional probability $P(L|Y)$, as $Y = y_1, \ldots, y_T$ is the per-frame output sequence and L is the target label sequence (Long et al. 2018), and train the classifier straight from the input sequences and map to the conditional probabilities of the possible outputs (Graves et al. 2008).

In the final step, for transcribing the recognized text characters sequence, the CTC layer is decoded to find the most plausible transcription (Yin et al. 2017). The transcribed sequences are transferred to a lexicon based decoding system in order to simulate the probable dependence along with the adjoining characters in the candidate words.

3.5 Defect Detection and Classification

Defect detection is pinpointing the location of defect in the pipe image (defect localization) and defect classification is to determine the type of detected defect. Introduced region based object detection frameworks in section 2.6 can be accomplished in three main steps: (i) descriptive region selection that scans the whole image with multi-scale sliding windows to identify all possible objects positions in the image; (ii) feature extraction to represent the features associated with different objects in the image and describe them; (iii) defect classification to identify the defect based on its distinguishable features from the other defects. These frameworks consist of several associated steps, and each of these steps including region proposal generation, feature extraction using deep neural networks, object classification, and bounding box regression required to be trained separately. So, the detection would be time consuming and computationally expensive. In this research, sewer defect detection has been conducted by the state-of-the-art object detection framework and their performance is compared to come up with the best defect detection model and its relative architecture. It is presented that one-shot models are performing better than two step models since the pixels are mapped directly to defect classes probabilities and bounding boxes coordinates.

The proposed defect detection framework is a customized single shot multibox detector (SSD) inspired from (Liu, Anguelov et al. 2015). The proposed framework starts with creating the dataset

for different defects and ground truth boxes for each defect. The framework training and evaluation steps are explained in the following.

3.5.1 Dataset

The dataset for training the defect detection framework is developed from the sewer images and reports introduced in section 3.2. the dataset needs to be annotated by labels for sewer defect types and a bounding box for the location of the defect in the images. In order to generate the required labeled dataset, the images were labeled manually using LabelImg graphical image annotation tool (Tzutalin 2015). In total 3500 defect images were labeled in three class of defects. The image labels and the bounding box coordinates were saved as XML files in PASCAL VOC format. The prepared annotation data needs some preprocessing and conversion to be applied as input for the proposed model.

3.5.2 Framework

The SSD framework consists of three main sections, feature extractor which is a pre-trained image classification architecture, auxiliary layers that map the higher level features into multi scale convolutional features, and prediction layers to classify and localize the objects in the image. In this research, the feature extractor layers are studied to find the best architecture to fit the sewer defects dataset. The auxiliary and prediction layers are inspired from the proposed framework by Liu et al. (2015) and customized for the problem in hand. In this section the steps of the framework are explained in detail. Figure 3-15 shows the architecture of the proposed SSD model.

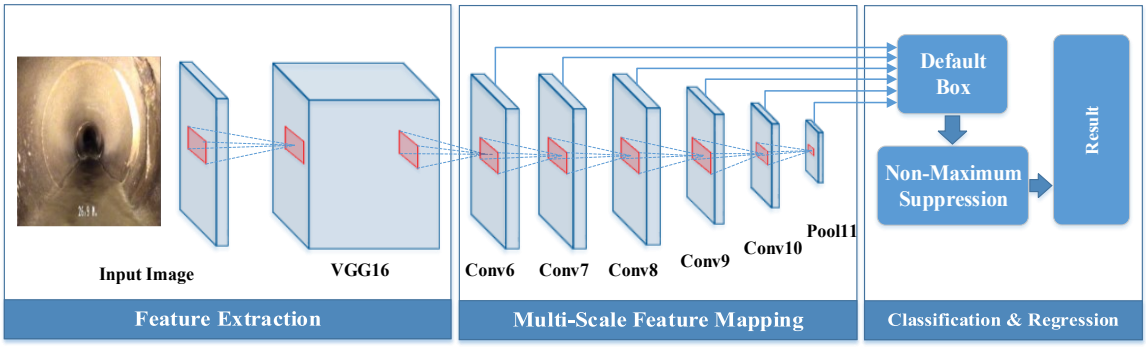


Figure 3-15. SSD architecture

In feature extractor part, the convolutional layers of an existing architecture are utilized to capture the low-level features of the image. These layers are pre-trained using transfer learning. In transfer learning the initial weights of the layers are borrowed from a closely related subject. This part can be replaced by other suitable architectures as it is discussed completely in the next chapter of this research. However, some modifications are supposed to be done. The input image is resized to 300 by 300 pixels. Also, in pooling layers mathematical functions as *ceiling* or *floor* are justified based on the feature maps dimension. The fully connected layers are removed from the end of the architecture since they are next to useless for the task.

The upcoming convolutional layers in the object detection framework, represent higher level features by combining the fed low-level features from the feature extractor layers. Six convolutional layers with various filter map sizes, construct the pyramid of the image features in

different scales. The extracted high level feature maps are fed to the prediction layers. Defects can happen in infinite positions and any possible form and scale. For defect prediction. Thereby, the search space would be required by a predetermined set of priors.

Priors are fixed boxes with predefined aspect ratios and positions of particular feature maps in the image. The priors are chosen and get matched sensibly with the ground truth bounding boxes of the dataset. They are positioned at all possible regions in all low level and high level feature maps of auxiliary layers. Priors are used in various scales regarding the feature map size and smaller scale priors are assigned to larger feature maps. Their scale starts from 0.1 to 0.9 of image dimensions. Moreover, a variety of prior aspect ratios are employed for every feature map. The prior ratios of 1:1, 2:1, and 1:2 for all feature maps, and two more 3:1 and 1:3 ratios for intermediate feature maps are used. In total, 8732 priors are specified for the feature maps of the auxiliary layers.

Prior is considered as a first guess for bounding box prediction. The coordinates of the priors are regressed with the coordinates of the ground truth bounding box. If the coordinates of prior would be (p_x, p_y, p_w, p_h) where p_x *and* p_y are coordinates of the prior center, and p_w *and* p_h are its width and height; and the bounding box coordinates would be (b_x, b_y, b_w, b_h) where b_x *and* b_y are coordinates of the box center, and b_w *and* b_h are its width and height, then the regress bounding box coordinates are [Equations 3-18 to 3-21]:

$$g_x = \frac{b_x - p_x}{p_w} \qquad \text{Equation 3-18}$$

$$g_y = \frac{b_y - p_y}{p_h} \qquad \text{Equation 3-19}$$

$$g_w = \log\left(\frac{b_w}{p_w}\right) \qquad \text{Equation 3-20}$$

$$g_h = \log\left(\frac{b_h}{p_h}\right) \qquad \text{Equation 3-21}$$

Therefore, for every prior at all feature map regions the regression of bounding box and the classes scores will be predicted. So, there will be 8732 predictions for regressions and class scores for all priors. The Jaccard index (Intersection-over-Union) is utilized to determine the overlap extents of the two boxes (Equation 3-22). In this research, a threshold of 0.6 is used for Jacquard index and priors with IoU less than the threshold are considered as no-object and the ones with equal and more than the threshold are positive matches. In result, each ground truth box can be matched with multiple overlapping default boxes.

$$IoU = \frac{A \cap B}{A \cup B} \qquad \text{Equation 3-22}$$

The training objective is to minimize the loss of two loss functions for confidence (L_{conf}) and bonding box location (L_{loc}). To calculate confidence loss function, the number of negative matches (i.e., background) needs to be restricted using hard negative mining. In hard negative mining the model only considers those predictions which were the hardest to recognize as no object. The

confidence loss is calculated as the sum of cross entropy losses between positive and no object predictions (Liu et al. 2015) (Equation 3-23):

$$L_{conf} = \frac{1}{n} \left(\sum_{positive} CE + \sum_{no\ object} CE\right) \quad \text{Equation 3-23}$$

The localization loss would be the is the smooth L1 loss between the predicted box (p) and the ground-truth box (g) (Girshick 2015) (Equation 3-24):

$$L_{loc} = \frac{1}{n} \sum_{positive} smooth_{L1} \quad \text{Equation 3-24}$$

So, the overall objective loss function is a weighted sum of these losses (Liu, Anguelov et al. 2015) (Equation 3-25):

$$L = L_{conf} + \alpha L_{loc} \quad \text{Equation 3-25}$$

When there are two or multiple boxes with positive prediction of the same object, they are considered redundant prediction. To avoid this problem Non-Maximum Suppression (NMS) method is used. First, the Jaccard index among all predicted boxes in a given class is calculated. If the index of two boxes is more than a defined threshold, then most likely these boxes are predicting one same object and the box with the maximum score will be kept and suppress the others.

3.6 Performance evaluation

Performance evaluation for the models represents the generalization capacity of them and their prediction accuracy. In this research, for each model of the proposed framework, a separate performance evaluation set up is introduced. The anomaly detection model is differentiating between a normal frame and anomalous frame, so a binary classifier performance evaluation system is proposed. The localization model is based on text detection and recognition algorithms. Thereby, evaluation metrics are introduced for text detection (i.e., text bounding box detection), and the accuracy of the text recognition model is calculated by comparing the test results and ground truth information. For the defect detection and classification model, the evaluation metrics for categorical classifiers are employed to assess the classifier model capabilities and detection performance.

3.6.1 Anomaly detection model

Based on (Olson and Delen 2008), if a video frame is fed to the anomaly detection model, there are four possible prediction outcomes. If the frame is an anomaly (positive) and it is detected as an anomaly (positive), it is considered to be a true positive (TP), and if the classifier predicts anomalous for a normal frame, it is counted as a false positive (FP). Correct classification of normal frames as negative would be regarded as a true negative (TN), and incorrect classification of the normal frame as an anomalous frame, would be counted as false negative (FN). The performance of the anomaly detector is evaluated by three measures introduced by (2008):

(1) Recall or sensitivity is calculated as the ratio of correctly classified positives (TP) and the total positive count (TP+FN).

$$\text{Recall (true positive rate)} = \frac{TP}{TP+FN} \qquad \text{Equation 3-26}$$

(2) Precision is calculated as the ratio of correctly classified positives (TP) and the total classified positives (TP+FP).

$$\text{Precision (sensitivity)} = \frac{TP}{TP+FP} \qquad \text{Equation 3-27}$$

(3) Overall prediction accuracy is calculated as the ratio of total correctly predicted positives and negatives (TP + TN) by the total number of examples (TP+TN+FP+FN).

$$\text{Accuracy} = \frac{TP+TN}{TP++TN+FN+FP} \qquad \text{Equation 3-28}$$

3.6.2 Frame localization

For text detection, various evaluation protocols were possible, including denoting the whole text blocks, words, and characters (Lucas 2005). In this research, the ability of the model to identify word rectangles in an image is measured since it is difficult to specify text blocks in distracted sewer images. For the text character recognition step the cropped word recognition capability is evaluated. Each of these protocols is explained in the next sections.

For text detection performance evaluation, the precision and recall metrics are used as recommended in Lucas (2005). Precision, *p* is the number of texts detected correctly divided by the total number of detections. The low precision score shows that the system overrates the number of text boxes. Recall, *r* is the number of texts detected correctly divided by the total number of targets. Low recall score shows that system underrates the number of text boxes. Also, it is unlikely the system behaves exactly like a human in detecting the bounding boxes for an identified word (Lucas 2005). So, for text bounding box matching, a flexible measure is defined as the area of intersection of two boxes divided by the area of the minimum bounding box containing both rectangles (Lucas 2005). So, two rectangles are considered matched when their intersection ratio is between 0.7 and 1. The best match m(r, R) for a box *r* in a set of boxes *R* is defined as (Lucas 2005):

$$m(r,R)=max\ mp(r,r')|r' \in R \qquad \text{Equation 3-29}$$

And, precision and recall are defined as:

$$Precision\ (p') = \frac{\sum re \in E\ m(re,T)}{|E|} \qquad \text{Equation 3-30}$$

$$Recall\ (r') = \frac{\sum rt \in T\ m(rt,E)}{|T|} \qquad \text{Equation 3-31}$$

Where *T* and *E* are the sets of targets and identified rectangles. An *f* metric is used to combine the estimated precision and recall into a single measure:

$$f = \frac{1}{\alpha/p\prime+(1-\alpha)/r\prime} \qquad \text{Equation 3-32}$$

Where controls the relative weights and α=0.5 defines equal weights for precision and recall. The performance of the text detection algorithm is evaluated on sewer images from the generated dataset and compared with manually labeled ground truth. The capability of the proposed model is measured by checking the correct detection of bounding boxes around text information in sewer images.

3.6.3 Defect detection and classification

In contrary to binary classification, in categorical classification and object detection, there are no true negatives (TN). True positive (TP) shows the number of real defects that are correctly detected as defects and false positive (FP) indicates the number of non-defected images that are predicted as defected. Meanwhile, the number of real defects which are detected as non-defect is a false negative (FN). So the precision and recall are calculated as:

$$\text{Recall (true positive rate)} = \frac{TP}{TP+FN} \qquad \text{Equation 3-33}$$

$$\text{Precision (sensitivity)} = \frac{TP}{TP+FP} \qquad \text{Equation 3-34}$$

Moreover, for defect detection in each bounding box, a confidence level, and related coordinates are determined. The overlap ratio of the predicted bounding box and the ground truth box can be used to determine TP and FP. the ratios above a certain value considered as TP. The ratio is calculated as (Everingham et al. 2012):

$$a_0 = \frac{extent(p \cap g)}{extent(p \cup g)} \qquad \text{Equation 3-35}$$

Where a_0 is the ratio of overlay among the predicted bounding box p and ground truth bounding box g. Besides calculating precision and recall, the area under the precision-recall curve is defined as Average Precision (AP) (Equation 3-31) (Everingham et al. 2012).

$$AP = \int_0^1 p(r)dr \qquad \text{Equation 3-36}$$

Mathematically, the precision value for recall ($\hat{r}$) is replaced with the maximum precision for any recall $\geq \hat{r}$ (Everingham et al. 2012).

$$p_{interp}(r) = \max p(\tilde{r}) \qquad \tilde{r} \geq r \qquad \text{Equation 3-37}$$

Eventually, *mAP* is calculated for all the target classes using Equation 3-33 (Everingham et al. 2012):

$$mAP = \frac{1}{N_t} \sum_i AP_i \qquad \text{Equation 3-38}$$

Where N_t is the number of target classes, and AP_i is the AP value for class *i*.

Chapter 4 : Model Implementation and Validation

In this chapter, the application of the proposed anomaly detection methodology described in chapter 3, is demonstrated in the context of a real-world example. The videos are taken from the sewer CCTV inspection videos of the City of Laval, Quebec, Canada. The material of sewer pipes in the data set were circular form concrete pipes, with 610 mm diameter. The video format is MPEG-2, at a frame rate of 30 frames per second and resolution of 640×480 pixels. The following section presents the implementation process of the proposed framework. Data is used to develop the model and verify the validation of the developed methodology.

4.1 Case study

Civic infrastructure inspection and assessment is a routine part of maintenance procedures in many municipalities all over the world. In this research, the inspection data has been provided by the City of Laval. The reports have been provided by SIMO Management Inc. where a camera recorded the inspection CCTV videos with a telephoto lens. The city of Laval intended to assess the condition of sewer pipelines and prioritize the sewer pipelines. The inspection project is a part of completing phase 2 of the City's action plan. The data has been used for both model development and testing, and the validation of the models.

The project, including inspection and assessment of 1900 manholes and sewer pipes in pre-defined areas in Laval. A total length of approximately 130 kilometers of sanitary and combined sewer pipes with different pipe diameters ranging from 150 mm to 1500 mm was analyzed and assessed to provide condition assessment of manholes and pipes. Figure 4-1 represents the areas where inspection has been conducted.

Figure 4-1. Inspected areas in Laval

Image from Simo Inc. report

Parts of the inspection video have been employed to develop the model. The data was divided into two parts, training data and testing data in order to evaluate the performance of the model. The training data also split up into videos including only normal frames, and videos including normal frames and frames containing anomalies in. Figure 4-2 shows the pipelines which the inspection videos are used for model development.

Figure 4-2. Part of inspected sewer pipelines
Image from Colmatec Inc. report

4.2 Anomaly detection model

In this section, the application of the proposed anomaly detection algorithm is demonstrated in the context of a real-world example. As mentioned before, the proposed model aims to perform anomaly detection in sewer pipelines inspection videos as an automated process. The capability and accuracy of developed models are tested using the datasets which are extracted from CCTV inspection videos obtained from the City of Laval, Quebec, Canada. The material of sewer pipes in the data set was concrete with a circular cross-section and 610 mm (24 inches) diameter. The video format is MPEG-2, at a frame rate of 30 frames per second and resolution of 640×480 pixels. The preprocessing of videos and SVM training have been done using MATLAB (MathWorks 2018). The

implementation of the models is described in the following steps. The data set split into a training set which includes only normal frames and a testing set which contains both normal and abnormal frames. The training subset contains 20 video samples, and the testing subset includes 12 video samples. Each sequence lasts around 1500 frames, for a total duration of 25 minutes.

4.2.1 Data preparation

The proposed approach is tested on sewer inspection videos from the mentioned data base. Sewer CCTV videos usually contain too many frames which do not contain any important information such as starting, under water, camera lens malfunctioning, and end of pipe frames. In this research, the mentioned frames have been neglected from being analyzed, and only forward view frames are captured. The data set split into a training set which includes only normal frames, and a testing set which contains both normal and abnormal frames. A preprocessing step applied to transform the frames to gray scale and resize them to 320 × 280 pixels. Moreover, due to the low speed of robot carrying the camera, sewer pipeline inspection videos contain a repeated scene with not much difference in consecutive frames. Thereby, to decrease the number of calculations, the frame rate reduced to 10 frames per seconds. Sewer inspection videos have traits that make them prone to noise because of sewer pipeline lighting condition and camera movements. Noise can be random or white noise, or coherent noise created by the device's mechanism or processing algorithms. Noise reduction step has been performed using a Gaussian filter of size [2, 2] with $\sigma = 1.1$. Figures 4-3 shows sample frames from the training sets with normal frames and anomalous frames, respectively.

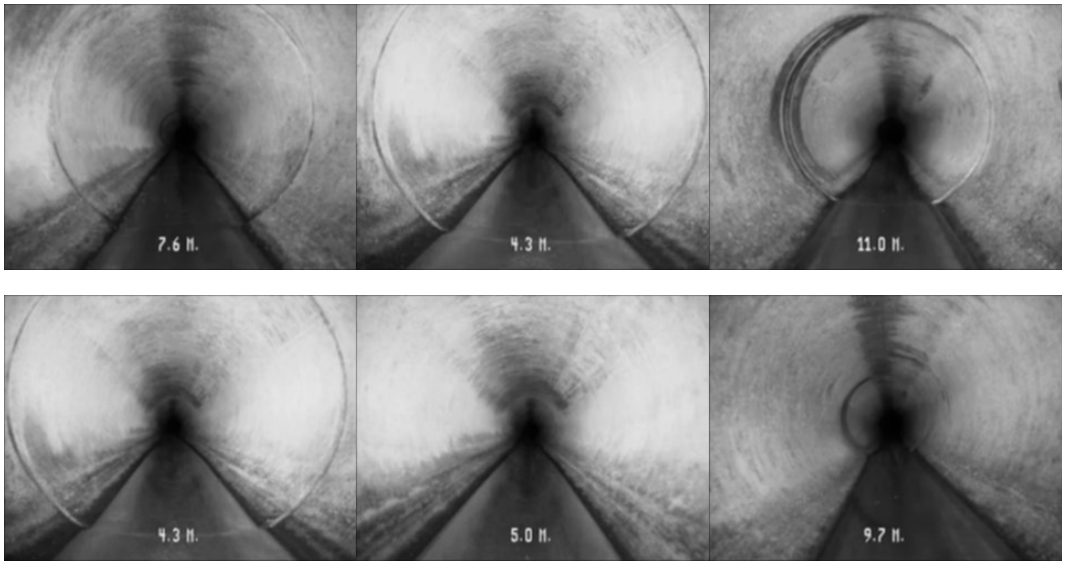


Figure 4-3. Sample of training set

4.2.2 Feature extraction and scene representation

In this step first, the preprocessing tasks are applied to the video. The frame rate is adjusted to 5 frames per seconds and then are converted from RGB to grayscale. The size of frames

is changed to 120 × 80 pixels, and a noise reduction step has been performed using a Gaussian filter of size [2,2] with σ = 1.1.

The prepared video frames then are fed to 3D SIFT feature sampling algorithm. The videos of the training dataset are analyzed, and the frames are patched by eight frames as cuboids temporal length, and each one is divided into eight sub-regions, two along each side. The interest points of training frames are extracted, and in all the sub-regions, the orientation is accumulated into an 8x4 histogram and configured as 4x4x4 sub-histograms as represented in figure 4-4. Thereby, the final descriptor of each key point is a 2048 (4×4×4 ×8×4) dimensions vector.

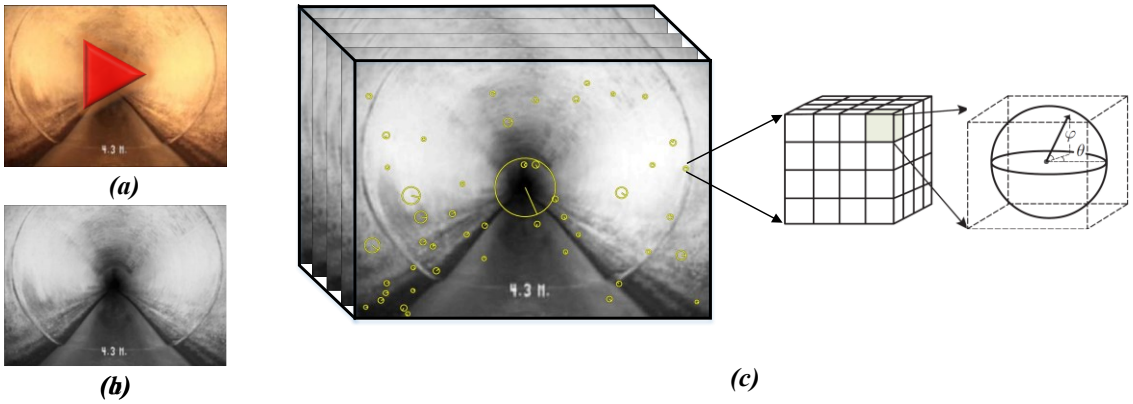


Figure 4-4. Illustrative example of 3D SIFT descriptor definition: (a) CCTV frame sequence; (b) grayscale and resized input frame; (c) construction of 3D sift descriptor (adapted from (Moradi et al. 2020)).

4.2.3 Training the SVM classifier and anomaly detection

After extracting the frames features, the data is mapped to a higher dimensional feature space to separate the data. For the data which is inseparable in the input space, the proper kernels are selected. In this paper, linear and nonlinear Gaussian radial basis function (RBF) kernels are used, and the performance of each is compared. In the next step, the appropriate parameters of a geometric figure are determined. This geometric figure is enclosing the feature space of sampled data vectors. In this research, the sampled data distribution is defined in a hyperplane with relative weight vector (w) and bias parameter (r). All data samples lie inside the hyperplane are considered as normal with relative Lagrange value of zero and the data with Lagrange value equal and greater than C are identified as outliers (Shahid et al. 2015).

The main part of the model training is the selection of proper values of γ and C. Large values of γ result in overtraining and too many support vectors, and on the other hand, small values of γ lead to few support vectors and affect the generalizing of the model. Also, the optimum value of C needs to be estimated to regulate the number of normal data that possibly will be classified as outliers in the model. Different values of C and γ in the range of 0.001 to 100,000, are examined to determine various combinations of sensitivity and providing the highest accuracy for the model. So, γ = 0.001 is considered as the initial γ

and is increased until there is no more decrease in support vectors. Eventually, the model showed the best performance by values of γ = 1.0 and *C*=10.by the accuracy of 0.95.

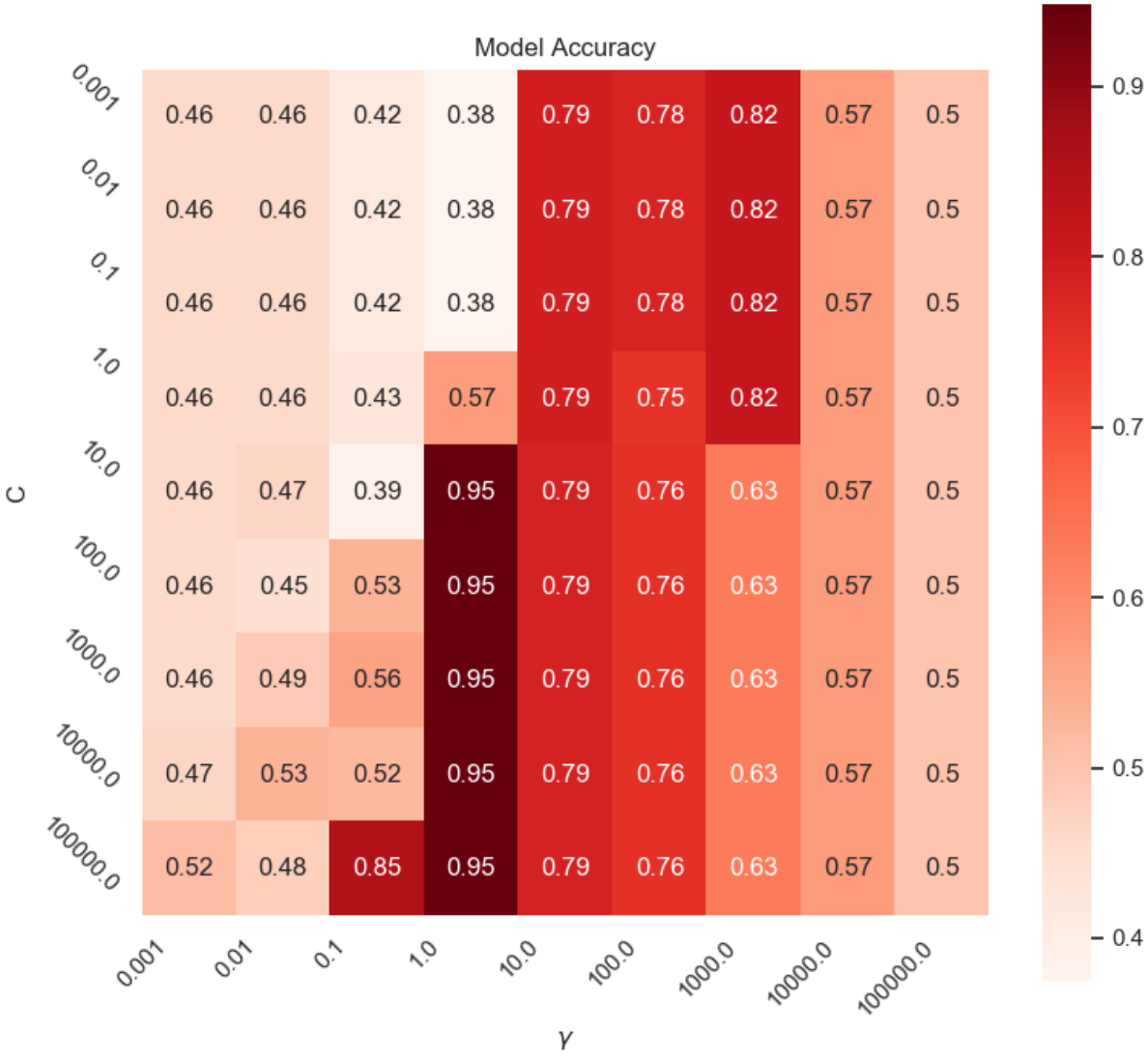


Figure 4-5. OC-SVM model accuracy by various C and γ

4.2.4 Performance evaluation

As mentioned in the last section, performance evaluation for classification models represents the generalization capacity of the classifier and its prediction accuracy. Based on (Olson and Delen 2008), given a classifier and an instance, there are four possible prediction outcomes as shown in Table 4-1.

Table 4-1. Elements of confusion matrix

	Predict (Anomaly)	Predict (Normal)
Actual (Anomaly)	TP	FN
Actual (Normal)	FP	TN

The performance of the proposed model is compared with three other models through the introduced metrics. A Multivariate Gaussian distribution-based anomaly detection model trained with SIFT descriptors extracted from the database. Multivariate Gaussian distribution assumes the normal data as Gaussian distributed (Do 2008). The prepared dataset is converted to a Gaussian distribution and feed to the model to fit and estimate two parameters of μ and Σ. The algorithm calculates the probability of data points using the following (Do 2008):

$$(p|\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{n}{2}}\,|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(d-\mu)^T\Sigma^{-1}(d-\mu)\right) \qquad \text{Equation 4-1}$$

Where μ is an n-dimensional vector, and Σ is an $n \times n$ covariance matrix.

When new data is given $p(d)$ would be computed and compared with ε as a predefined threshold. If $p(d) < \varepsilon$ then the data point is identified as an anomaly, otherwise it is considered normal.

Moreover, two other OC-SVM models trained with SIFT and GIST descriptors. GIST descriptors rely on the shapes in an image and condenses the gradient information of various regions of the image to provide a global description of the scene (Douze et al. 2009). GIST descriptor is computed by convolving the input image with 32 Gabor filters at four scales and eight orientations, generating 32 feature maps with a matching size of the input image. Then, every feature map is divided into 16 regions in 4×4 grids, and in each region of interest feature values are averaged. The resulted averaged values from the 16 regions included in 32 feature maps are concatenated to turn out a 512 features GIST descriptor (Zhang 2015). Also, both trained models used RBF kernels.

The trained models were tested against a data set with 300 anomalous (positive) frames and 300 normal (negative) frames, which were previously unseen by the models. The resulted performance metrics, including recall, precision, accuracy, and F1 score are presented as a confusion matrix in Table 4-2. The recall which represents the fraction of frames are correctly labeled as anomalous out of all anomalous frames in the dataset. The proposed model outperforms the other models by recall rate of 0.93 which shows the capability of the proposed model in recognize the anomalies correctly. The number of false alarms (false positive) has been reduced but still is high in the proposed model due to the low quality of sewer images. Also, higher recall is more desirable for the problem in hand since it shows the model does not miss anomalous frames among all the frames. All the OC-SVM models show almost the same precision rate of 0.80-0.82 which illustrates the fraction of frames are correctly labeled as anomalous (i.e., true positives) out of all the frames that the classifier labeled as anomalous. Moreover, regarding the accuracy, the proposed model outperforms other models by accuracy of 86.67%. However, the high

accuracy does not approve the model's ability because it deals both false positive and false negetive equally. Therefore, F1 Score is calculated for all the models for better evaluation. Among all four tested models, the proposed model performs better than the others by F1 Score of 0.88.

Table 4-2. Prediction performance metrics of the proposed model through testing data sets

	Precision	Recall	Accuracy	F1 Score
Multivariate Gassian-D- SIFT	0.62	0.58	61.23%	0.60
OC-SVM- GIST	0.80	0.54	70.55%	0.65
OC-SVM- SIFT	0.81	0.67	75.33%	0.77
OC-SVM – 3D SIFT	0.82	0.93	86.67%	0.88

Receiver operating characteristic (ROC) curve represent the different values of classifier recognition rate corresponding to various false positive rates. The area under the ROC curve (AUC) should tend to 1 to show the prediction ability of the classifier and AUC is less than 0.5 shows that the classifier recall is only rested on probability (Shahid et al. 2015). As presented in figure 4-6, the AUC of the trained model with 3D SIFT features and RBF kernel was found to be 0.966. The results reveal that it is feasible to use the proposed approach for automated sewer defect detection in CCTV videos. However, accuracy can be improved by reducing false alarms, which mostly are because of sudden changes in camera angle or water level. Overall, evaluation results show that the proposed model is suited for identification and localization of anomalies in sewer CCTV inspection videos since that it employs temporal information from sequences of frames rather than single static frame image.

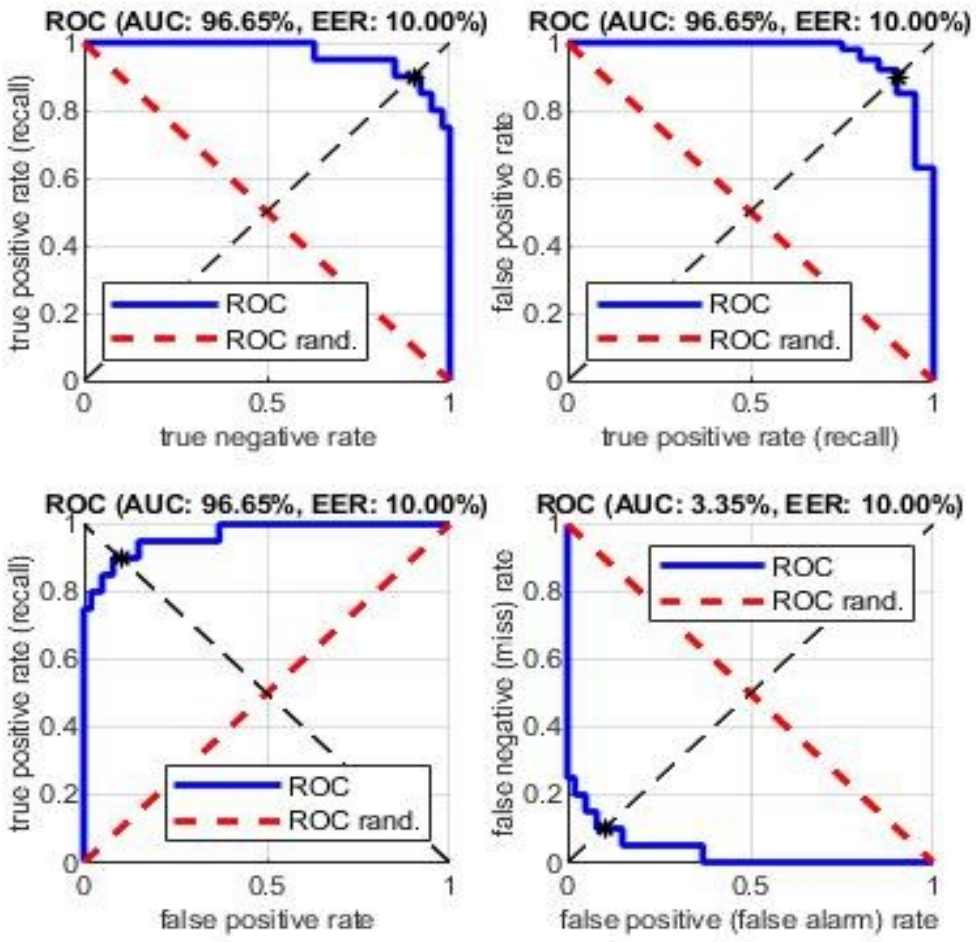


Figure 4-6. The receiver operating curves (ROC) for OC-SVM on the testing dataset

4.3 Frame Localization

The performance of proposed text detection and text recognition are evaluated on the provided dataset. The videos only included text information for frame location, which is shown as travel distance from the manhole. However, the proposed system can detect and recognize other text information such as sewer pipe location, pipe section, date, and defects information which is registered by the operator and can be used for off-site evaluation and quality control purposes.

4.3.1 Data preparation

The text information in sewer video frames include English alphabetical and numerical characters. So, to provide the training dataset for the text recognition model, one single dataset was generated for all types of text information from cropped texts. The dataset size was boosted using data augmentation technique and image transformations including scaling, color channels shifting, adding various noises, and rotation (Figure 4-7).

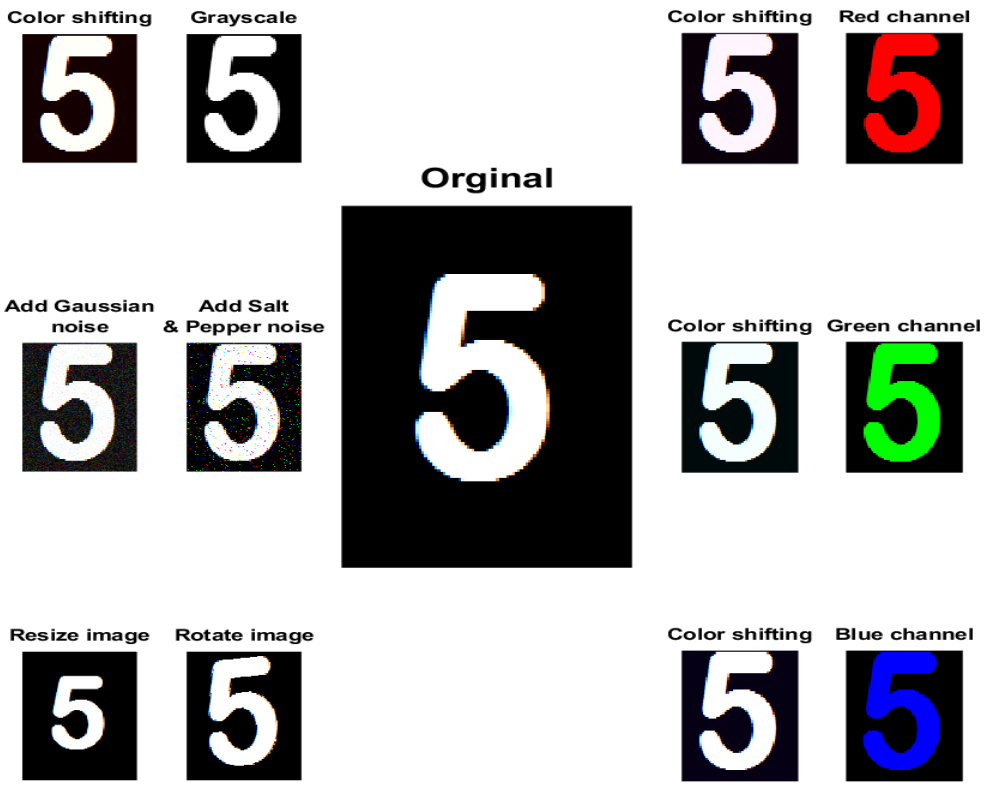


Figure 4-7. Data augmentation using various transformations (adapted from (Moradi et al. 2020))

4.3.2 Text detection

The text detection approach introduced in section 3.4.1 is utilized to detect the text information in separated anomalous frames. The quality of the images was enhanced by sharpening the edges and noise reduction. The extracted MSER regions were identified and various filters were applied to exclude non-text regions and the relative text boxes were highlighted. The highlighted texts in the inspection video frames were only the distance from the starting manhole in meters.

The performance of the text detection was assessed by the introduced metrics in section 3.6.2. The precision and recall metrics were calculated by studying the total number of correct detected texts and the ground truth bounding boxes. The low precision score indicates that the text detector overestimates the amount of text bounding boxes. On the other hand, low recall score illustrates that the text detector underestimates the number of text bounding boxes. Table 4-3 shows the evaluated precision, recall, and f_{score} of the text detection model.

Table 4-3.Text detection algorithm evaluation results

	Precision	Recall	f_{score}
Text detector	0.73	0.60	0.66

4.3.3 Text recognition

The highlighted text boxes were fed to the text recognition model to classify the included characters. The text recognition model was trained on a synthetic 62- way dataset which was generated from different types of texts in sewer inspection video frames. The trained character classifier achieved the accuracy of 86.6% on recognition of the characters. Table 4-4 presents the accuracies of developed text recognition model versus optical character recognition (OCR) model. The OCR model was created using MATLAB OCR trainer toolbox (2018). the proposed text recognition model shows better performance on the sewer images comparing the OCR model since "the OCR algorithms performing well on clean images and sharped text edges while sewer images are too noisy, and usually the texts are blurred" (Moradi et al. 2020).

Table 4-4. Cropped word recognition results (adapted from (Saeed Moradi et al., 2020))

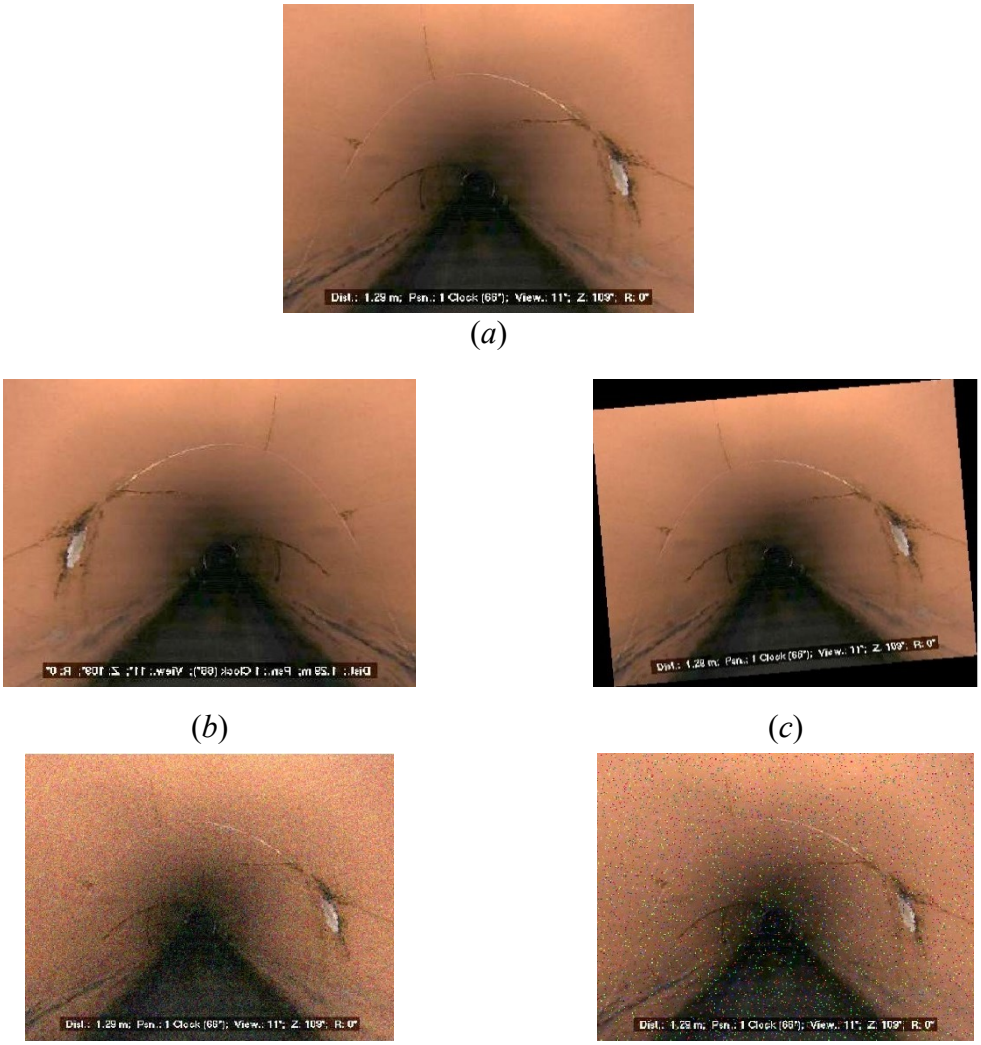
Method	Accuracy
Proposed	86.6%
OCR	57%

4.4 Defect detection and classification

It is believed that the foremost causes of sewer pipeline incidents include pipe blockages, which are mainly caused by defects such as deposits, the disproportion of inflow and outflow caused by infiltration, and pipe wall breakages which can be triggered by cracks (EPA 2004). Thus, the prepared dataset includes four types of defects including joint displacement, deposit, infiltration, and crack.

4.4.1 Dataset preparation

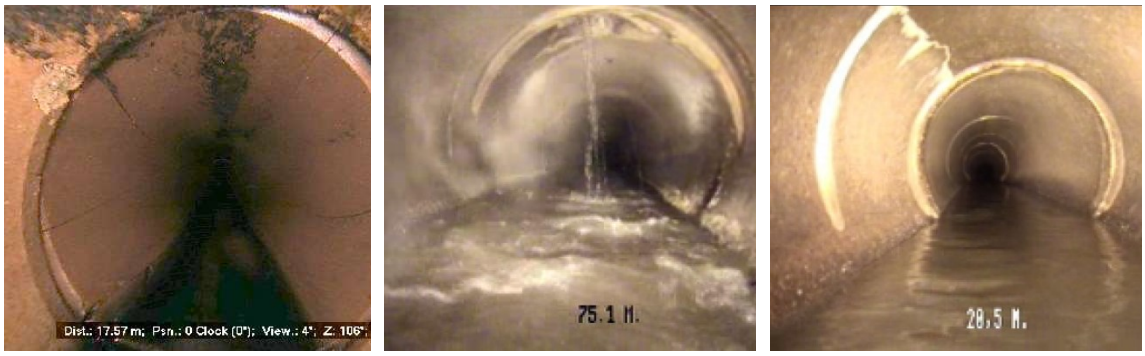
The collected images are augmented by different transformations and adding noise. Images are rotated by 180 degrees and flipped to transform the location of the defects in images. Moreover, noise filters applied to add Gaussian noise to the histogram of the images and salt and pepper noise to the image pixels. After data augmentation, the size of dataset increased considerably and a total of 6000 images used for training the model. Figure 4-8 shows an example of data augmentation operations used in the experiment.



(*a*) (*b*) (*c*) (*d*) (*e*)

Figure 4-8. Image augmentation: (a) Original image; (b) Horizontal flip; (c) Image rotation; (d) Gaussian noise; (e) Salt & Pepper noise.

To label the dataset, both the defect type and its location in the images should be annotated. The dataset is prepared for four sewer defects, including crack, deposit, infiltration, and joint displacement images. For each category, the different subtypes are ignored, and the category name is used as a target label for all related images. So, various types of cracks, including longitudinal cracks, diagonal cracks, and complex cracks, are labeled as crack. In the same way, different types of deposits such as attached deposits and settled deposits are labeled as deposit. Images containing multiple defects are also included in the dataset. Figure 4-9 shows images with cracks and joint displacement, deposit, and infiltration, and crack and infiltration.



(*a*) (*b*) (*c*)

Figure 4-9.example of images with multiple defects: (*a*) cracks and joint displacement; (*b*) deposit and infiltration; (*c*) crack and infiltration

Although images with higher resolution can provide more information, the training computational cost increases significantly. On the other hand, for getting more reliable results, the input images in the testing dataset should have the same size as training and validation datasets. Therefore, for training defect detection and classification model, all the images are resized to 300×300 pixels. All images are annotated to introduce ground truth bounding boxes and target labels using LabelImage (Tzutalin 2015) graphical annotation tool. The XML files generated by LabelImage need to be converted to TensorFlow records to be used by the developed models. Figure 4-10 shows LabelImage graphical interface.

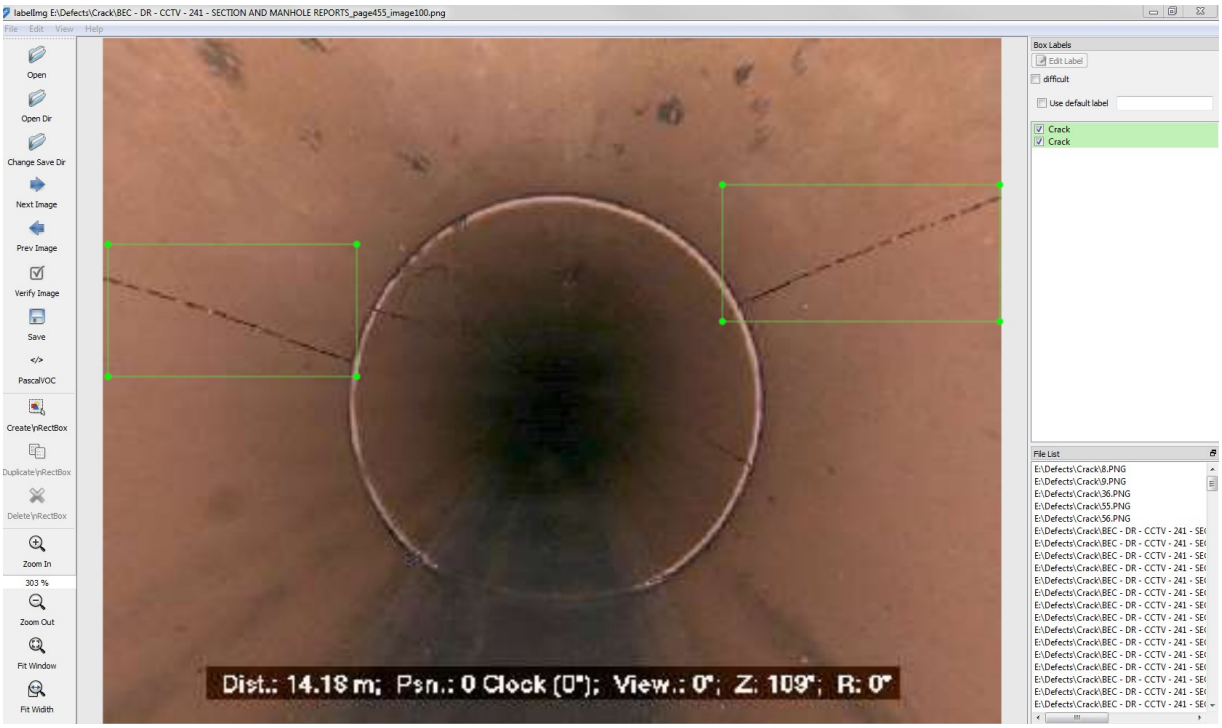


Figure 4-10. LabelImage image annotation

4.4.2 Experiments and results

Several experiments are carried out to examine the performance of the proposed framework and other frameworks in the detection of different defects in the prepared dataset and also the influence of pre-training network type and hyper parameters on model performance. In each experiment, 70% of the dataset is used as a training set, 10% as a validation set, and 20% as the testing set. Also, all the experimental models are developed using Keras (Chollet 2015) high-level API with TensorFlow (Martin et al. 2015) backend, which provides libraries to create various layers of deep learning architecture. The models were run on the same machine with Windows operating system with an Intel Core i7-4790 CPU, two Nvidia GeForce GTX 1070 GPU and 32G Ram.

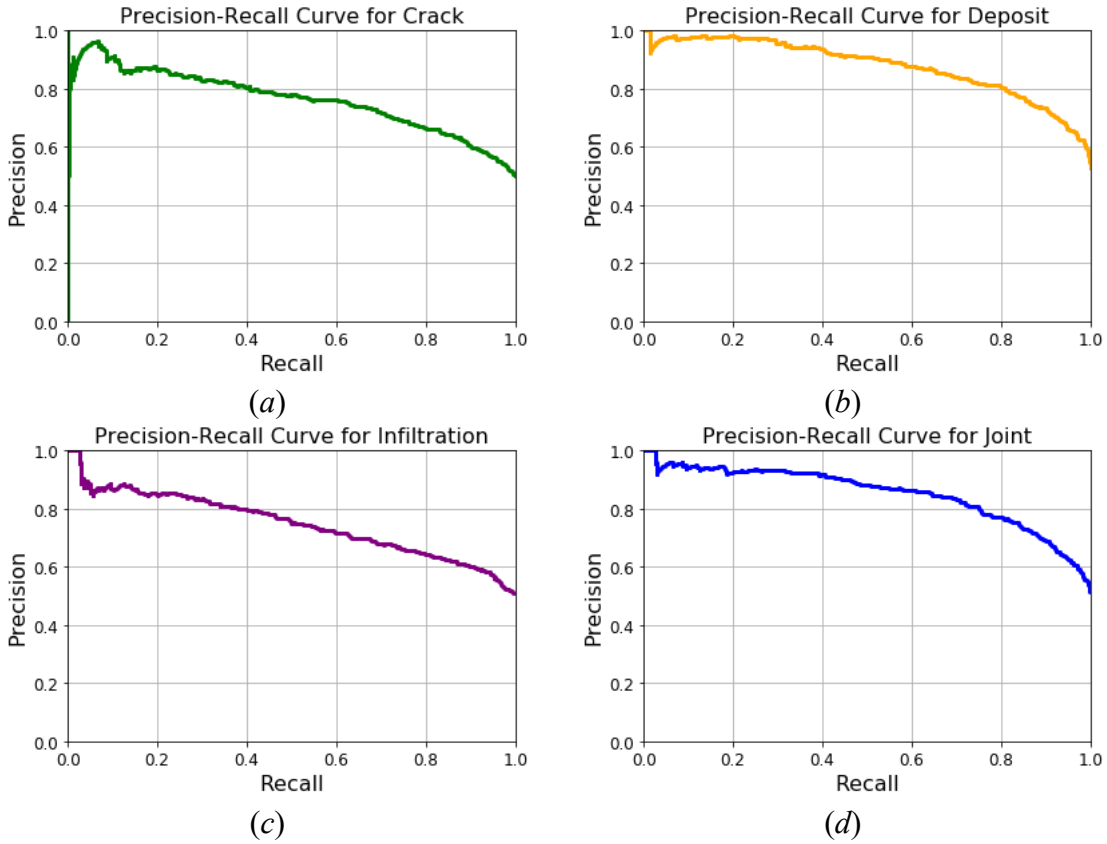
Experiment 1

In the first experiment, the ability of the model in detecting and classifying each of the four defects is evaluated. The precision and recall are calculated for each of them in the base model trained by VGG16 as initializer and five convolutional layers for feature extraction and mapping default boxes. The number of images samples for all defects are the same, and for each of the defects, 3500 images were used for training and test the model. The proposed model is evaluated by *AP* of each class and *mAP* of the model. Table 4-5 shows the *AP* and *mAP* of the model.

Table 4-5. mAP and AP of a model for different defects (%)

Model	Pre-training	mAP	AP (%)			
			Crack	Deposit	Infiltration	Joint Displacement
SSD300	VGG16	79.6	76.3	88.2	74.9	81.3

The model shows better performance for more distinguishable defects such as deposit and joint displacement. However, in defects such as cracks and infiltration, the AP results are slowly less as of 76.3% and 74.9% respectively. Potential reasons can be color resemblance, geometry of the defects, and intensity changes among the features. Also, fine-grained nature of these type of objects makes a big challenge for the predictor model to distinguish them accurately. In addition, image noise and illumination affect the accuracy of the model, particularly in dealing with these types of defects.



(*a*) (*b*)

(*c*) (*d*)

Figure 4-11. Precision-recall curve: (*a*) Crack; (*b*) Deposit; (*c*) Infiltration; (*d*) Joint displacement

The precision-recall curve for each defect is calculated with various confidence thresholds (Figure 4-11). Obviously, by increasing the confidence threshold, the number of FNs predictions decreases, and on the other hand, the number of FPs increases. Therefore, in precision-recall curves, the precision value drops as recall value increases in each model prediction step. In sewer defect detection, it is far more important not to miss the possible defects, so higher recall rate among the predictions is more crucial. Therefore, lower FNs prediction is more desired which results in that the model will not miss the potential pipe defects.

Experiment 2

In another experiment, the performance of five of the most common object detection frameworks including R-CNN (Girshick et al. 2014), Fast R-CNN (Girshick 2015), Faster R-CNN (Ren et al. 2017), YOLO (Redmon et al. 2016), and SSD (Liu et al. 2015) were compared. The frameworks used initialization weights trained on ImageNet classification dataset (Russakovsky et al. 2015) and the object detection models are trained on the prepared dataset of sewer defects. The defect detection results are compared based on the evaluated *mAP*. The assessed frameworks Table 4-6shows comparative results of different frameworks trained and tested on the provided dataset.

Table 4-6. Comparative results of different object detection frameworks

Framework	Learning Method	AP				mAP
		Crack	Deposit	Infiltration	Joint Displacement	
R-CNN	SGD.BP	68.1	71.8	45.3	69.8	63.8
Fast R-CNN	SGD	77.0	78.4	59.6	82.6	74.4
Faster R-CNN	SGD	84.3	82.0	67.8	88.6	80.7
YOLO	SGD	77.4	77.0	43.3	85.3	70.8
SSD	SGD	76.3	88.2	74.9	85.7	81.3

As illustrated in the table, among the tested frameworks SSD model outperformed the other frameworks in sewer defect detection. The first two frameworks, R-CNN, and Fast R-CNN use selective search method for region proposals and achieved *mAP* of 63.8% and 74.4%. Meanwhile, Faster R-CNN uses RPN, which still is time-consuming in training and detection, and the resulted *mAP* is 80.7%. However, regression models such as YOLO and SSD are faster at the cost of a decrease in prediction accuracy and achieved *mAP* of 70.1% and 81.3%. Since in sewer defect detection, the aim is to inspect in real time, regression-based models are preferred. Thereby, the SDD object detection framework has been selected due to its better performance comparing to YOLO framework.

Experiment 3

In the next experiment, various pre-trained models are examined as initialization network for feature extractor in SSD object detection approach to find the most proper one in sewer defect detection and classification. The reviewed models are top rated models in ILSVRC ImageNet (Russakovsky et al. 2015) including AlexNet (Krizhevsky et al. 2012), VGGNet (Simonyan and Zisserman 2014b), GoogleNet (Szegedy et al. 2014), and ResNet (Kaiming et al. 2015). The models are pre-trained on MS-COCO (Lin et al. 2014). The models are fine-tuned and trained six convolutional layers in SSD framework using the provided dataset of sewer defect images.

All four trained models are evaluated by evaluation dataset by three criteria: classification loss, localization loss, and total loss of the model. In classification both models trained by GoogleNet and VGGNet showed better accuracy of 92.8% and 92.10% respectively, comparing to ResNet with accuracy of 91.68% and AlexNet with accuracy of 89.9%. Figure 4-12 shows the comparative loss of four different models.

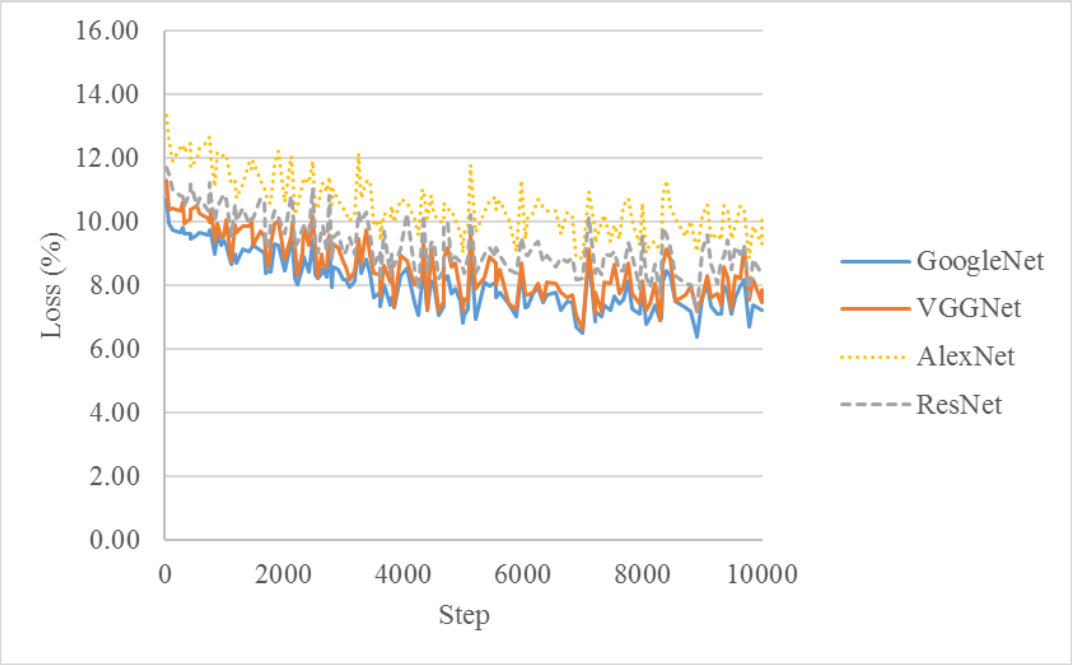


Figure 4-12. Classification Loss with different pre-training models

Regarding the localization loss, GoogleNet showed a considerable accuracy of 94.32%, which is higher the accuracy of other models. The pre-trained model with VGGNet achieved the accuracy of 93.60% while ResNet and Alex net achieved the accuracy of 92.63% and 91.88% respectively. Figure 4-13 represents localization loss of different models.

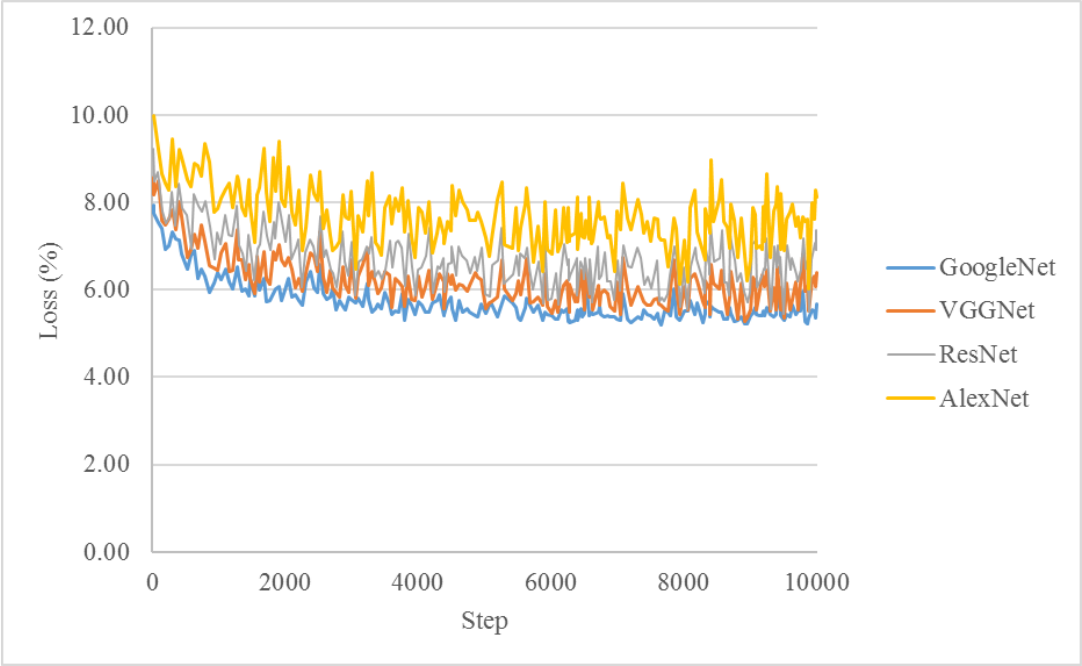


Figure 4-13. Localization Loss with different pre-training models

In a total loss, the model with pre-trained GoogleNet network achieved the accuracy of 91.44% while the other models achieved 88.69%, 86.85%, and 85.18% accuracy by VGGNet, ResNet, and AlexNEt respectively. Therefore, GoogleNet is selected as initialization network for defect detection model. Figure 4-14 describes the error rate of different models with various training networks.

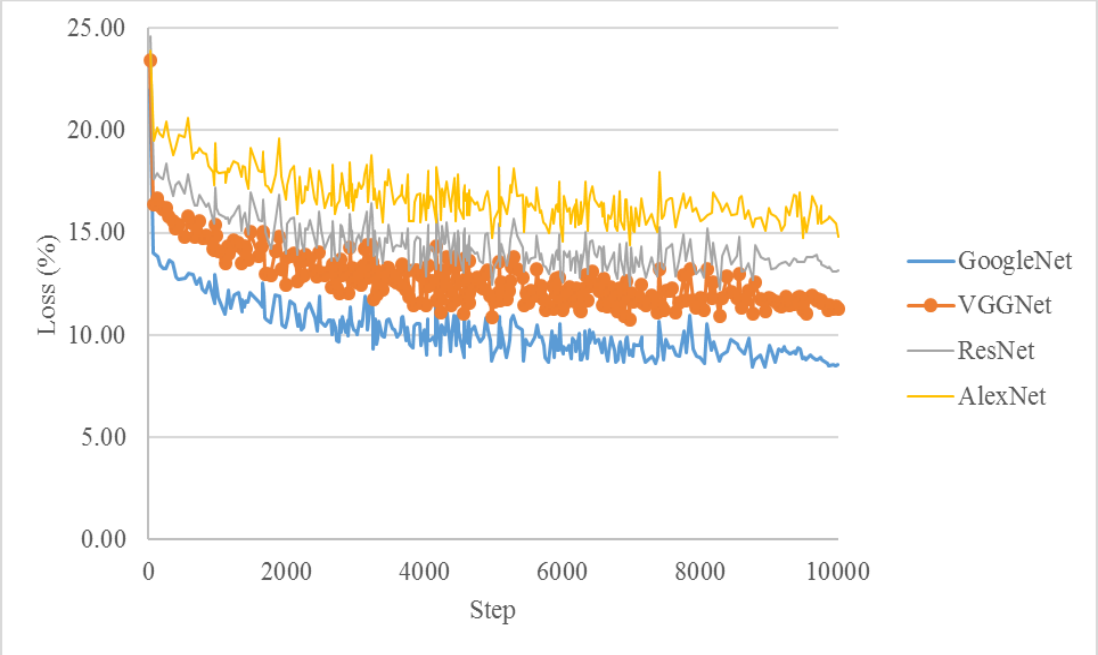


Figure 4-14. Total Loss with different pre-training models

4.4.3 Model validation

The proposed framework for sewer defect detection is selected and trained based on the results of the previous experiments. In the first experiment, the ability of different object detection frameworks in sewer defect detection is evaluated, and the SSD framework presented the best performance comparing the other frameworks. So, in the second experiment, various pre-trained networks are examined as the backbone for the SSD object detection framework, and GoogleNet showed higher performance comparing the other networks. Therefore, the final model is developed based on SSD framework with GoogleNet backbone.

The proposed defect detection and classification model was validated with the prepared test dataset. The test data fed into the developed model to detect and recognize the defects and validated with the provided ground truth. The results showed acceptable performance in practice for defect detection where the accuracy of 84.4% is achieved. The figures 4-15 to 4-17 show the detection results in sewer pipeline images.

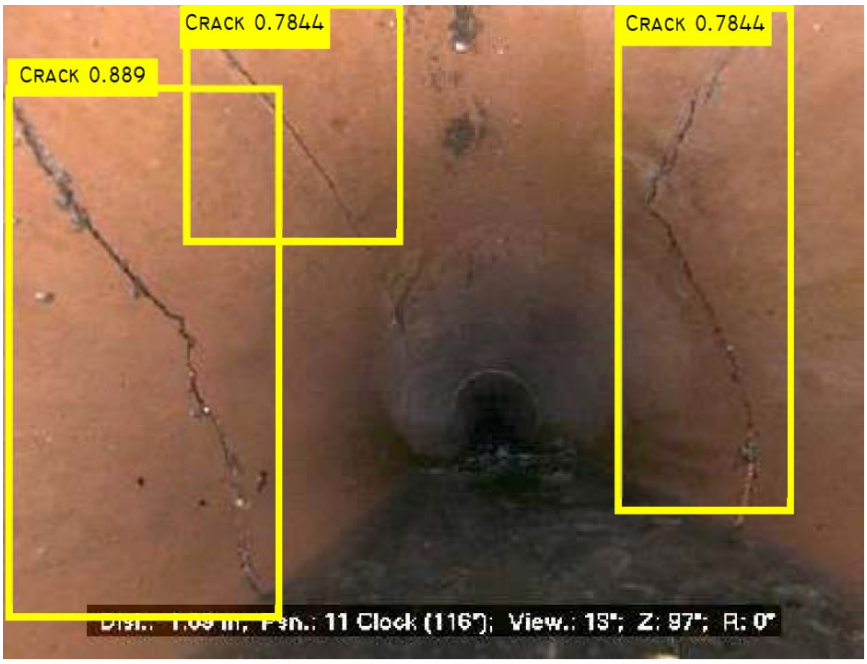


Figure 4-15. Example image with multiple cracks

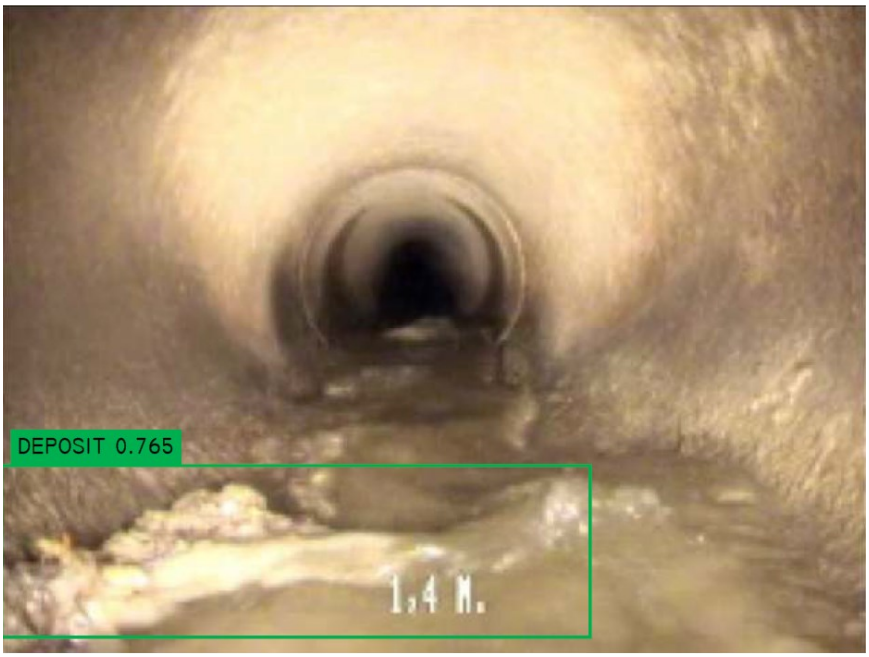


Figure 4-16. Example image with deposit

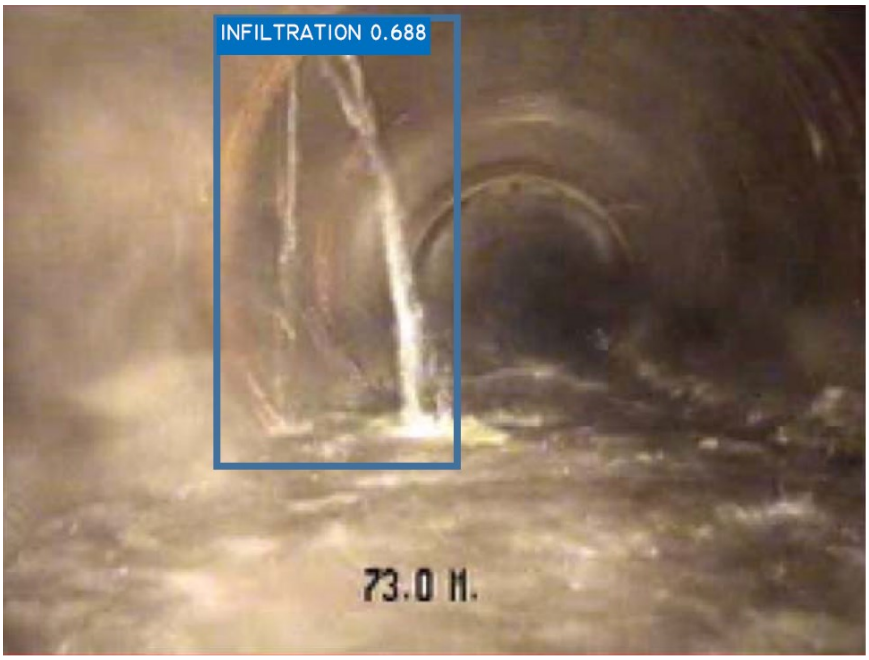


Figure 4-17. Example of image with infiltration

The proposed approach is employed on static images. However, the framework cab be justified to detect the defects in inspection videos as well. Regarding the availability of large inspections video data, the proposed framework would be helpful to analyze them. Moreover, there have been a few incorrect predictions for some defects. Probable reason can be the higher level features resemblance among the confused defects. Also, cluttered background in sewer images is a big obstruction for correct detection. Larger training datasets with the variety in defect conditions and backgrounds may improve the prediction performance of the framework.

Chapter 5 : Conclusions, Contributions and Future Work

In this chapter, the different contributions of this research are presented in addition to the conclusion and limitations of the proposed models. Also, several suggestions are presented to enhance the frameworks performance and the potential areas for improvement in sewer pipeline assessment automation. The suggestions are applicable for future studies in the area and visual inspection automation studies in other underground infrastructure.

5.1 Summary

The main objective of this research was to develop an automated CCTV inspection tool for sewer pipelines. To achieve the main objective, two sub-models were developed. First, a novel approach for anomaly detection in sewer pipeline inspection videos has been proposed. There are almost infinite patterns for each sewer defect and using algorithms like pattern recognition and change detection seems not to be efficient. The trained OC-SVM based model is proposed to deal with real-world observations and numerous feasible patterns of anomalies in sewer pipelines. The model uses 3D SIFT features to model scene dynamics and appearance information. The approach is composed of demonstrating conditions reflected as normal and distinguishing outliers to them. Moreover, the approach would be able to conduct real-time detection and localization of anomalies in sewer inspection videos. In the following steps, the identified frames were localized using the text information included in sewer inspection video frames. So, the frame location in the pipe segment would be extracted and notified.

In the second sub-model, a deep learning based approach was proposed to detect and classify defects among the identified anomalous frames. After comparing various object detection frameworks, SSD framework was selected as the base model for the object detection task. The framework was customized for the problem in hand and trained using collected data sets from CCTV inspection videos. The capability of the proposed framework defect detection and classification in anomalous frames was validated through different experiments. Moreover, the proposed framework was modified by tuning different hyper-parameters and the parameters of layers were justified to study the most influential factors on the performance. The influence of initialization networks was tested using state-of-the-art networks and the achieved localization, classification, and total accuracies were compared. It was depicted that the networks with deeper convolutional layers such as GoogleNet can improve the performance of the model. Although much deeper layers can extract more detailed and accurate image features, the computation time will increase exponentially. So, the best balance of accuracy and computational cost was achieved by the proposed framework.

It is supposed that the automated inspection tool would help municipalities and practitioners to overcome their main problems in sewer inspection by reducing subjectivity and increasing productivity of condition assessment job. The application of the proposed

framework has been illustrated by a simple real-world CCTV inspection video of a sewer pipeline.

5.2 Concluding remarks

- An anomaly detection model was developed using an innovative spatio-temporal feature capturing and training an OC-SVM to classify the frames into two classes as normal and anomalous frames.
- The anomaly detection model was tested, and the model achieved a recall rate of 0.973, precision rate of 0.941, and accuracy of 0.956.
- The frames were located by a novel end to end text detection and recognition model to extract the location information in inspection video frames. To overcome the specific pipeline images condition, various image processing techniques were evaluated. MSER method was used to detect the probable text regions, and non-text regions were filtered using CC labeling criteria and stroke width calculations. Then, the text in detected textboxes was recognized by a CNN to classify the characters and predict the transcriptions.
- The text detector model achieved the recall and precision rates of 0.73 and 0.60 respectively and f_{score} of 0.60. The accuracy of text recognition model reached to 0.866 in the evaluation by test dataset.
- The identified anomalous frames are fed to the developed deep object detection model, which is fine-tuned using transfer learning and trained by the provided dataset. Batch normalization and drop out techniques were used to avoid overfitting. Also, hyperparameters such as convolutional filters dimensions and their strides were justified to increase the accuracy of the model. In the training phase, the warm-up algorithm was used to schedule the learning rate in various training epochs.
- The accuracy of defect detection and classification model was increased from 81% in the first developed network to 91.44% after applying the proposed adjustments.

5.3 Contributions

This research proposed a two-step approach for sewer CCTV inspection automation, which provides references for practitioners to apply the proposed computer vision and deep learning techniques to address similar problems in infrastructure visual inspection. The practical application of the proposed approach is expected to make a considerable reduction in inspection time and cost, as well as to improve the accuracy of sewer pipeline assessment. The contributions of this research can be summarized as:

- A comprehensive introduction and comparison of various sewer in section technologies.
- A through literature review of the research works on sewer inspection automation.
- Automated anomaly detection through sewer CCTV inspection videos and localizing them in the sewer pipe segment.
- Automated defect detection and classification in sewer CCTV inspection videos.

The following section presents the limitations of this research and potential areas of enhancement.

5.4 Limitations

It is obvious that the pattern recognition and object detection is still an active area in computer vision science, and more research is required to adopt the new techniques to sewer pipeline assessment methods. The proposed method for feature extraction requires improvements in sampling part and using clustering tools. Also, anomaly detection model can be justified with various kernel tricks to make them more generalized and increase the prediction accuracy.

The current proposed defect detection model was tested on still images, and in case of applying the model on inspection videos, the network needs to be modified to reach real-time frame rate speed. Also, there were several wrong predictions in some defects such as cracks and infiltration, and the model got confused in distinguishing among these defects. The potential reasons can be similarity in geometrical shape, same color intensity, and pattern changes. However, more study on the model's architecture is required to increase the performance of the deep network in the classification of the images taken under different environmental conditions.

5.5 Recommendations and Future Research

For sewer pipeline defect detection, the proposed models can be enhanced, and the research can be extended by providing the following.

5.5.1 Models enhancement

- For anomaly detection model, the emerging techniques in machine learning area can be employed. The author thinks that deep learning algorithms such as auto encoders can be proper tools due to the availability of more powerful computational hardware resources.
- Frame localization can be revised to increase the accuracy by considering the video frames timeline and calculating the frame location based on tractor speed and video frame rate.
- The defect detection frameworks can be improved by employing applicable algorithms from other areas of deep learning like Natural Language Processing (NLP). The architecture can be modified using concepts such as attention models for classification performance improvement or transformers for transfer learning.
- The defect detection and classification framework can be modified for real time detection in inspection videos. The model’s hyper parameters need to be justified to be applicable on videos.
- The defect detection can be expanded to detect different types of each defect such as various types of crack.
- The accuracy of developed models can be increased by better input data as well as more input data.

5.5.2 Recommendation for Future Work

- Bigger datasets can be provided for each defect with the collaboration of industries and governmental agencies. Also, a standard data set can be prepared to be introduced as a benchmark for future academic research.
- An automated Graphical User Interface (GUI) can be designed to facilitate the application of defect detection models for the inspectors and end users.
- Employing emerging computer vision methods to quantify the detected defects and determine the severity of the defects to use in defect specific assessments.
- An assessment model can be developed to integrate with the defect detection model to estimate the pipeline and network indices.
- A decision making model can be developed using expert systems to correlate with defect detection and assessment models.

Chapter 6 References

Adams, J. (2010). *Side-Scan Sewer Inspection Comes of Age*. New Jersey, USA.

ASCE. (2017). *Report Card for America's Infrastructure*. Reston, Virginia, United States.

Chae, M., and Abraham, D. (2001). "Neuro-Fuzzy Approaches for Sanitary Sewer Pipeline Condition Assessment." *Journal of Computing in Civil Engineering*, 15(1), 4–14.

Chae, M., Iseley Tom, and Abraham Dulcy M. (2003). "Computerized Sewer Pipe Condition Assessment." *New Pipeline Technologies, Security, and Safety*, Proceedings, ASCE, 477–493.

Chaki, A., and Chattopadhyay, T. (2010). "An intelligent fuzzy multifactor based decision support system for crack detection of underground sewer pipelines." *2010 10th International Conference on Intelligent Systems Design and Applications*, 1471–1475.

Chen, H., Tsai, S. S., Schroth, G., Chen, D. M., Grzeszczuk, R., and Girod, B. (2011). "Robust text detection in natural images with edge-enhanced Maximally Stable Extremal Regions." 2609–2612.

Chen, K., Hu, H., Chen, C., Chen, L., and He, C. (2018). "An Intelligent Sewer Defect Detection Method Based on Convolutional Neural Network." *2018 IEEE International Conference on Information and Automation (ICIA)*, 1301–1306.

Cheng, J. C. P., and Wang, M. (2018). "Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques." *Automation in Construction*, 95, 155–171.

Chollet, F. (2015). "Keras."

Cortes, C., and Vapnik, V. (1995). "Support-Vector Networks." *Machine Learning*, 20(3), 273–297.

Costello, S. B., Chapman, D. N., Rogers, C. D. F., and Metje, N. (2007). "Underground asset location and condition assessment technologies." *Tunnelling and Underground Space Technology*, 22(5–6), 524–542.

Dang, L. M., Hassan, S. I., Im, S., Mehmood, I., and Moon, H. (2018). "Utilizing text recognition for the defects extraction in sewers CCTV inspection videos." *Computers in Industry*, 99, 96–109.

Dirksen, J., Clemens, F. H. L. R., Korving, H., Cherqui, F., Le Gauffre, P., Ertl, T., Plihal, H., Müller, K., and Snaterse, C. T. M. (2013). "The consistency of visual sewer inspection data." *Structure and Infrastructure Engineering*, 9(3), 214–228.

Do, C. B. (2008). "The multivariate gaussian distribution." *Section Notes, Lecture on Machine Learning, CS 229*.

Douze, M., Jégou, H., Harsimrat, S., Amsaleg, L., and Schmid, C. (2009). “Evaluation of GIST descriptors for web-scale image search.” pp.19:1–8.

ECT Team, P. (2007). *Sewer Scanner and Evaluation Technology (SSET)*. USA.

EPA. (2004). *Report to congress on impacts and control of combined sewer overflows and sanitary sewer overflows*. Washington, D.C.

Epshtein, B., Ofek, E., and Wexler, Y. (2010). “Detecting text in natural scenes with stroke width transform.” 2963–2970.

Erfani, S. M., Rajasegarar, S., Karunasekera, S., and Leckie, C. (2016). “High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning.” *Pattern Recognition*, 58, 121–134.

Everingham, M., VanGool, L., Williams, C. K. I., Winn, J., and Zisserman, A. (2012). “The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.”

Fang, X., Guo, W., Li, Q., Zhu, J., Chen, Z., Yu, J., Zhou, B., and Yang, H. (2020). “Sewer Pipeline Fault Identification Using Anomaly Detection Algorithms on Video Sequences.” *IEEE Access*, 8, 39574–39586.

FCM. (2019). *Canadian Infrastructure Report Card (CIRC)*. Ottawa, ON,Canada.

Fukushima, K. (1975). “Cognitron: A self-organizing multilayered neural network.” *Biological Cybernetics*, Springer-Verlag, 20(3–4), 121–136.

Gao, W., Zhang, X., Yang, L., and Liu, H. (2010). “An improved Sobel edge detection.” 67–71.

Girshick, R. (2015). “Fast R-CNN.” *CoRR*, abs/1504.08083.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). “Rich feature hierarchies for accurate object detection and semantic segmentation.” IEEE, 580–587.

González, Á., Bergasa, L. M., Yebes, J. J., and Bronte, S. (2012). “Text location in complex images.” 617–620.

Gonzalez, R. C., and Woods, R. E. (2006). *Digital Image Processing*. Prentice-Hall, Inc, Upper Saddle River, NJ, USA.

Graves, A., Liwicki, M., Horst, B., Schmidhuber, J., and Santiago, F. (2008). “Unconstrained On-line Handwriting Recognition with Recurrent Neural Networks.” *Advances in Neural Information Processing Systems 20*, J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, eds., Curran Associates, Inc, 577–584.

Guo, W., Soibelman, L., and Garrett, J. (2007). “Automatic Defect Detection and Recognition for Asset Condition Assessment: A Case Study on Sewer Pipeline Infrastructure System.” ASCE, 419–426.

Guo, W., Soibelman, L., and Garrett, J. (2009a). “Automated defect detection for sewer

pipeline inspection and condition assessment." *Automation in Construction*, 18(5), 587–596.

Guo, W., Soibelman, L., and Garrett, J. (2009b). "Visual Pattern Recognition Supporting Defect Reporting and Condition Assessment of Wastewater Collection Systems." *Journal of Computing in Civil Engineering*, 23(3), 160–169.

Halfawy, M., and Hengmeechai, J. (2014a). "Efficient Algorithm for Crack Detection in Sewer Images from Closed-Circuit Television Inspections." *Journal of Infrastructure Systems*, 20(2), 4013014.

Halfawy, M. R., and Hengmeechai, J. (2014b). "Optical flow techniques for estimation of camera motion parameters in sewer closed circuit television inspection videos." *Automation in Construction*, 38(Supplement C), 39–45.

Halfawy, M. R., and Hengmeechai, J. (2014c). "Automated defect detection in sewer closed circuit television images using histograms of oriented gradients and support vector machine." *Automation in Construction*, 38, 1–13.

Hao, T., Rogers, C. D. F., Metje, N., Chapman, D. N., Muggleton, J. M., Foo, K. Y., Wang, P., Pennock, S. R., Atkins, P. R., Swingler, S. G., Parker, J., Costello, S. B., Burrow, M. P. N., Anspach, J. H., Armitage, R. J., Cohn, A. G., Goddard, K., Lewin, P. L., Orlando, G., Redfern, M. A., Royal, A. C. D., and Saul, A. J. (2012). "Condition assessment of the buried utility service infrastructure." *Tunnelling and Underground Space Technology*, 28, 331–344.

Hawari, A., Alamin, M., Alkadour, F., Elmasry, M., and Zayed, T. (2018). "Automated defect detection tool for closed circuit television (cctv) inspected sewer pipelines." *Automation in Construction*, 89, 99–109.

Iyer, S., and Sinha, S. K. (2005). "A robust approach for automatic detection and segmentation of cracks in underground pipeline images." *Image and Vision Computing*, 23(10), 921–933.

Iyer, S., and Sinha, S. K. (2006). "Segmentation of Pipe Images for Crack Detection in Buried Sewers." 21(6), 395–410.

Jahne, B. (2002). *Digital Image Processing*. Springer-Verlag New York, Inc, Secaucus, NJ, USA.

Jain, A. K. (1989). *Fundamentals of Digital Image Processing*. Prentice-Hall, Inc, Upper Saddle River, NJ, USA.

Joachims, T. (1998). *Making Large-scale SVM Learning Practical. , LS VIII-Report*. Dortmund.

Kaiming, H., Xiangyu, Z., Shaoqing, R., and Jian, S. (2015). "Deep Residual Learning for Image Recognition." *CoRR*, abs/1512.0.

Kirstein, S., Müller, K., Walecki-Mingers, M., and Deserno, T. M. (2012). "Robust

adaptive flow line detection in sewer pipes." *Automation in Construction*, 21, 24–31.

Koch, C., Georgieva, K., Kasireddy, V., Akinci, B., and Fieguth, P. (2015). "A review on computer vision based defect detection and condition assessment of concrete and asphalt civil infrastructure." *Advanced Engineering Informatics*, 29(2), 196–210.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). "ImageNet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds., Curran Associates, Inc, 1097–1105.

Kumar, S. S., and Abraham, D. M. (n.d.). "A Deep Learning Based Automated Structural Defect Detection System for Sewer Pipelines." *Computing in Civil Engineering 2019*, 226–233.

Kumar, S. S., Abraham, D. M., Jahanshahi, M. R., Iseley, T., and Starr, J. (2018). "Automated defect classification in sewer closed circuit television inspections using deep convolutional neural networks." *Automation in Construction*, 91, 273–283.

Kumar, S. S., Mingzhu, W., Abraham, D. M., Jahanshahi, M. R., Tom, I., and Cheng, J. C. (2020). "Deep Learning–Based Automated Detection of Sewer Defects in CCTV Videos." *Journal of Computing in Civil Engineering*, 34(1), 4019047.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). "Deep learning." *Nature*, 521, 436.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). "Gradient-based learning applied to document recognition." *Proceedings of the IEEE*, 86(11), 2278–2324.

Lee, J., Lee, P., Lee, S., Yuille, A., and Koch, C. (2011). "AdaBoost for Text Detection in Natural Scene." 429–434.

Li, D., Cong, A., and Guo, S. (2019). "Sewer damage detection from imbalanced CCTV inspection data using deep convolutional neural networks with hierarchical classification." *Automation in Construction*, 101, 199–208.

Li, Y., and Lu, H. (2012). "Scene text detection via stroke width." 681–684.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Doll\'ar, P., and Zitnick, C. L. (2014). "Microsoft COCO: Common Objects in Context." D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds., Springer International Publishing, 740–755.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. E., Fu, C.-Y., and Berg, A. C. (2015). "SSD: Single Shot MultiBox Detector." *CoRR*, abs/1512.02325.

Long, S., He, X., and Ya, C. (2018). "Scene Text Detection and Recognition: The Deep Learning Era." *CoRR*, abs/1811.04256.

Lowe, D. G. (1999). "Object recognition from local scale-invariant features." *Proceedings of the IEEE International Conference on Computer Vision*, IEEE, 1150–1157.

Lowe, D. G. (2004). “Distinctive Image Features from Scale-Invariant Keypoints.” *International Journal of Computer Vision*, 60(2), 91–110.

Lucas, S. (2005). “ICDAR 2005 text locating competition results.” IEEE, 80–84 Vol. 1.

Makar, J. (1999). “Diagnostic Techniques for Sewer Systems.” *Journal of Infrastructure Systems*, 5(2), 69–78.

Martel, K., Feeney, C., and Tuccillo, M. (2011). *Field Demonstration of Condition Assessment Technologies for Wastewater Collection Systems*. Washington, DC, USA.

Martin, A., Ashish, A., Paul, B., Eugene, B., Zhifeng, C., Craig, C., Greg, S. C., Andy, D., Jeffrey, D., Matthieu, D., Sanjay, G., Ian, G., Andrew, H., Geoffrey, I., MichaelI, sard, Jia, Y., Rafal, J., Lukasz, K., Manjunath, K., Josh, L., Dandelion, M., Rajat, M., Sherry, M., Derek, M., Geoffrey, I., Mike, S., Jonathon, S., Benoit, S., Ilya, S., Kunal, T., Paul, T., Vincent, V., Vijay, V., Fernanda, V., Oriol, V., Pete, W., Martin, W., Martin, W., Yuan, Y., and Xiaoqiang, Z. (2015). “TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems.”

Mashford, J., Rahilly, M., Davis, P., and Burn, S. (2010). “A morphological approach to pipe image interpretation based on segmentation by support vector machine.” *Automation in Construction*, 19(7), 875–883.

Matas, J., Chum, O., Urban, M., and Pajdla, T. (2004). “Robust wide-baseline stereo from maximally stable extremal regions.” *Image and Vision Computing*, 22(10), 761–767.

MathWorks, I. (2018). “MATLAB Version 9.4.0 R2018a.”

McKim, R. A., and Sinha, S. K. (1999). “Condition assessment of underground sewer pipes using a modified digital image processing paradigm.” *Tunnelling and Underground Space Technology*, 14, 29–37.

Meijer, D., Scholten, L., Clemens, F., and Knobbe, A. (2019). “A defect classification methodology for sewer image sets with convolutional neural networks.” *Automation in Construction*, 104, 281–298.

Mnih, V., Heess, N., and Graves, A. (2014). “Recurrent models of visual attention.” Curran Associates, Inc., 2204–2212.

Mohamed, E., Tarek, Z., and Alaa, H. (2019). “Multi-Objective Optimization Model for Inspection Scheduling of Sewer Pipelines.” *Journal of Construction Engineering and Management*, 145(2), 4018129.

Moradi, S., and Zayed, T. (2017). “Real-Time Defect Detection in Sewer Closed Circuit Television Inspection Videos.” Proceedings, 295–307.

Moradi, S., Zayed, T., and Golkhoo, F. (2018a). “Automated sewer pipeline inspection using computer vision techniques.” *Pipelines 2018: Condition Assessment, Construction, and Rehabilitation - Proceedings of Sessions of the Pipelines 2018 Conference*.

Moradi, S., Zayed, T., and Golkhoo, F. (2018b). "Automated Sewer Pipeline Inspection Using Computer Vision Techniques." Proceedings, ASCE, 582 – 587.

Moradi, S., Zayed, T., and Golkhoo, F. (2019a). "Review on Computer Aided Sewer Pipeline Defect Detection and Condition Assessment." *Infrastructures*, 4(1), 10.

Moradi, S., Zayed, T., and Golkhoo, F. (2019b). "Review on computer aided sewer pipeline defect detection and condition assessment." *Infrastructures*, 4(1).

Moradi, S., Zayed, T., and Hawari, A. H. (2016). "Automated detection of anomalies in sewer closed circuit television videos using proportional data modeling."

Moradi, S., Zayed, T., Nasiri, F., and Golkhoo, F. (2019c). "Application of Deep Neural Networks in Sewer Pipeline Defect Detection and Classification." *No Dig 2019*, No Dig, Calgary, AB, 1–12.

Moradi, S., Zayed, T., Nasiri, F., and Golkhoo, F. (2020). "Automated Anomaly Detection and Localization in Sewer Inspection Videos Using Proportional Data Modeling and Deep Learning–Based Text Recognition." *Journal of Infrastructure Systems*, 26(3), 4020018.

Moselhi, O., and Shehab-Eldeen, T. (1999). "Automated detection of surface defects in water and sewer pipes." *Automation in Construction*, 8(5), 581–588.

Moselhi, O., and Shehab-Eldeen, T. (2000). "Classification of Defects in Sewer Pipes Using Neural Networks." *Journal of Infrastructure Systems*, 6(3), 97–104.

Myrans, J., Everson, R., and Kapelan, Z. (2018). "Automated detection of faults in sewers using CCTV image sequences." *Automation in Construction*, 95, 64–71.

Myrans, J., Everson, R., and Kapelan, Z. (2019). "Automated detection of fault types in CCTV sewer surveys." *Journal of Hydroinformatics*, 21(1), 153–163.

Najafi, M. (2016). *Pipeline infrastructure renewal and asset management*. McGraw-Hill Education: New York, Chicago, San Francisco, Athens, London, Madrid, Mexico City, Milan, New Delhi, Singapore, Sydney, Toronto.

NASSCO. (2001). *Pipeline Assessment & Certification Program (PACP) Reference Manual*. Marriottsville, MD.

Ni, D., Chui, Y. P., Qu, Y., Yang, X., Qin, J., Wong, T.-T., Ho, S. S. H., and Heng, P. A. (2009). "Reconstruction of volumetric ultrasound panorama based on improved 3D SIFT." *Computerized Medical Imaging and Graphics*, 33(7), 559–566.

O'Keefe, A. (2013). "Comprehensive Sewer Condition Assessment Using CCTV and Electro Scan: International Cases." ASCE, 113–123.

Olson, D. L., and Delen, D. (2008). *Advanced Data Mining Techniques*. Springer Publishing Company, Incorporated, Verlag Berlin Heidelberg.

Opitz, M., Diem, M., Fiel, S., Kleber, F., and Sablatnig, R. (2014). "End-to-End Text

Recognition Using Local Ternary Patterns, MSER and Deep Convolutional Nets." 186–190.

Pan, X., Clarke, T. A., and Ellis, T. J. (1994). "Detection and tracking of pipe joints in noisy images." *Videometrics III*, International Society for Optics and Photonics, 136–147.

Pan, Y., Hou, X., and Liu, C. (2009). "Text Localization in Natural Scene Images Based on Conditional Random Field." 6–10.

Plihal, H., Kretschmer, F., Ali, M. T. Bin, See, C. H., Romanova, A., Horoshenkov, K. V, and Ertl, T. (2016). "A novel method for rapid inspection of sewer networks: combining acoustic and optical means." *Urban Water Journal*, 13(1), 3–14.

Qidwai, U., and Chen, C. H. (2009). *Digital Image Processing: An Algorithmic Approach with MATLAB*. Chapman & Hall CRC, USA.

Rahman, S., and Vanier, D. (2004). *An Evaluation of Condition Assessment Protocols for Sewer Management*. Canada.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). "You Only Look Once: Unified, Real-Time Object Detection." 280–292.

Ren, S., He, K., Girshick, R., and Sun, J. (2017). "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6), 1137–1149.

Reyna, S., Vanegas, J., and Khan, A. (1994). "Construction Technologies for Sewer Rehabilitation." *Journal of Construction Engineering and Management*, 120(3), 467–487.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). "ImageNet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision (IJCV)*, 115(3), 211–252.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). "Estimating the Support of a High-Dimensional Distribution." *Neural computation*, 13(7), 1443–1471.

Scovanner, P., Ali, S., and Shah, M. (2007). "A 3-dimensional Sift Descriptor and Its Application to Action Recognition." MM '07, ACM, 357–360.

Selvakumar, A., Tuccillo, M., Martel, K., Matthews, J., and Feeney, C. (2014). "Demonstration and Evaluation of State-of-the-Art Wastewater Collection Systems Condition Assessment Technologies." *Journal of Pipeline Systems Engineering and Practice*, 5(2), 4013018.

Shahid, N., Naqvi, I. H., and Qaisar, S. Bin. (2015). "One-class support vector machines: analysis of outlier detection for wireless sensor networks in harsh environments."

Artificial Intelligence Review, 43(4), 515–563.

Shehab, T., and Moselhi, O. (2005). "Automated Detection and Classification of Infiltration in Sewer Pipes." *Journal of Infrastructure Systems*, 11(3), 165–171.

Simonyan, K., and Zisserman, A. (2014a). "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556*.

Simonyan, K., and Zisserman, A. (2014b). "Very Deep Convolutional Networks for Large-Scale Image Recognition." *CoRR*, abs/1409.1556, n.page.

Sinha, S. K., and Fieguth, P. W. (2006a). "Neuro-fuzzy network for the classification of buried pipe defects." *Automation in Construction*, 15(1), 73–83.

Sinha, S. K., and Fieguth, P. W. (2006b). "Segmentation of buried concrete pipe images." *Automation in Construction*, 15(1), 47–57.

Sinha, S. K., Fieguth, P. W., and Polak, M. A. (2003). "Computer Vision Techniques for Automatic Structural Assessment of Underground Pipes." 18(2), 95–112.

Sinha, S. K., and Karray, F. (2002). "Classification of underground pipe scanned images using feature extraction and neuro-fuzzy algorithm." *IEEE Transactions on Neural Networks*, 13(2), 393–401.

Sinha, S. K., Karray, F., and Fieguth, P. W. (1999). "Underground pipe cracks classification using image analysis and neuro-fuzzy algorithm." *Proceedings of the 1999 IEEE International Symposium on Intelligent Control Intelligent Systems and Semiotics (Cat. No.99CH37014)*, 399–404.

Sinha, S. K., and Knight, M. A. (2004). "Intelligent System for Condition Monitoring of Underground Pipelines." *Computer-Aided Civil and Infrastructure Engineering*, 19(1), 42–53.

Sinha Sunil K. (2001). "Automated Condition Assessment of Buried Sewer Pipeline Using Computer Vision Techniques." *Pipelines 2001*, Proceedings, ASCE, San Diego, California, United States, 1–12.

del Solar, J., and Köppen, M. (1996). "Sewage pipe image segmentation using a neural based architecture." *Pattern Recognition Letters*, Neural Networks for Computer Vision Applications, 17(4), 363–368.

Sonka, M., Hlavac, V., and Boyle, R. (2007). *Image Processing, Analysis, and Machine Vision*. Thomson-Engineering, USA.

Statistics Canada. (2018). *Canada's Core Public Infrastructure Survey: Wastewater and solid waste assets*. Ottawa, ON.

Su, T.-C., and Yang, M.-D. (2014). "Application of Morphological Segmentation to Leaking Defect Detection in Sewer Pipelines." *Sensors*, 14(5), 8686–8704.

Su, T.-C., Yang, M.-D., Wu, T.-C., and Lin, J.-Y. (2011). "Morphological segmentation

based on edge detection for sewer pipe defects on CCTV images." *Expert Systems with Applications*, 38(10), 13094–13114.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2014). "Going Deeper with Convolutions." *CoRR*, abs/1409.4842.

Tuccillo, M. E., Holley, J., Martel, K., and Boyd, G. (2010). *Report on condition assessment technology of wastewater collection systems*. USA.

Uijlings, J. R. R., van de Sande, K. E. A., Gevers, T., and Smeulders, A. W. M. (2013). "Selective Search for Object Recognition." *International Journal of Computer Vision*, 104(2), 154–171.

Viola, P., and Jones, M. (2001). "Rapid object detection using a boosted cascade of simple features." IEEE, 511.

Wang, M. Z., and Cheng, J. C. P. (2019). "Semantic Segmentation of Sewer Pipe Defects Using Deep Dilated Convolutional Neural Network." *ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction; Waterloo*, IAARC Publications, Waterloo, Canada, Waterloo, 586–594.

Wirahadikusumah, R., Abraham, D. M., Iseley, T., and Prasanth, R. K. (1998). "Assessment technologies for sewer system rehabilitation." *Automation in Construction*, 7(4), 259–270.

Wu, W., Liu, Z., and He, Y. (2015). "Classification of defects with ensemble methods in the automated visual inspection of sewer pipes." *Pattern Analysis and Applications*, 18(2), 263–276.

Xie, Q., Li, D., Xu, J., Yu, Z., and Wang, J. (2019). "Automatic Detection and Classification of Sewer Defects via Hierarchical Deep Learning." *IEEE Transactions on Automation Science and Engineering*, 16(4).

Yang, M.-D., and Su, T.-C. (2008). "Automated diagnosis of sewer pipe defects based on machine learning approaches." *Expert Systems with Applications*, 35(3), 1327–1337.

Yang, M.-D., and Su, T.-C. (2009). "Segmenting ideal morphologies of sewer pipe defects on CCTV images for automated diagnosis." *Expert Systems with Applications*, 36(2, Part 2), 3562–3573.

YANG, M.-D., SU, T.-C., PAN, N.-F., and LIU, P. E. I. (2011). "No Title." *International Journal of Wavelets, Multiresolution and Information Processing*, 09(02), 211–225.

Yang, M., Rajasegarar, S., Erfani, S. M., and Leckie, C. (2019). "Deep Learning and One-class SVM based Anomalous Crowd Detection." *Proceedings of the International Joint Conference on Neural Networks*, Institute of Electrical and Electronics Engineers Inc.

Ye, X., Zuo, J., Li, R., Wang, Y., Gan, L., Yu, Z., and Hu, X. (2019). "Diagnosis of sewer

pipe defects on image recognition of multi-features and support vector machine in a southern Chinese city." *Frontiers of Environmental Science & Engineering*, 13(2), 17.

Yin, F., Wu, Y.-C., Zhang, X.-Y., and Liu, C.-L. (2017). "Scene Text Recognition with Sliding Convolutional Character Models." *CoRR*, abs/1709.01727.

Yin, S., Zhu, X., and Jing, C. (2014). "Fault detection based on a robust one class support vector machine." *Neurocomputing*, 145, 263–268.

Yin, X., Chen, Y., Bouferguene, A., Zaman, H., Al-Hussein, M., and Kurach, L. (2020). "A deep learning-based framework for an automated defect detection system for sewer pipes." *Automation in Construction*, 109, 102967.

Zhang, C. (2015). "Computer Vision: What is a GIST descriptor?"

Zhao, J. Q., McDonald, S. E., and Kleiner, Y. (2001). *Guidelines for Condition Assessment and Rehabilitation of Large Sewers*. Canada, Ottawa.

Zhao, Z. Q., Zheng, P., Xu, S. T., and Wu, X. (2019). "Object Detection with Deep Learning: A Review." *IEEE Transactions on Neural Networks and Learning Systems*, 30(11), 3212–3232.