Obstruction Level Detection of Sewers Videos Using Convolutional Neural Networks

Mario A. Gutiérrez-Mondragón
Computer Science Department, Universitat Politecnica de Catalunya (UPC)-Barcelona Supercomputing Center (BSC)
Barcelona, Spain
Email: mario.alberto.gutierrez@upc.edu

Dario Garcia-Gasulla and Sergio Alvarez-Napagao
Barcelona Supercomputing Center (BSC)/High Performance Artificial Intelligence, Barcelona, Spain
Email: {dario.garcia, sergio.alvarez}@bsc.es

Jaume Brossa-Ordoñez and Rafael Gimenez-Esteban
Water Technology Center, Barcelona, Spain
Email: {jaume.brossa, rafael.gimenez}@cetaqua.es

***Abstract*—Worldwide, sewer networks are designed to transport wastewater to a centralized treatment plant to be treated and returned to the environment. This is a critical process for preventing waterborne illnesses, providing safe drinking water and enhancing general sanitation in society. To keep a perfectly operational sewer network several inspections are manually performed by a *Closed-Circuit Television* system to report the obstruction level which may trigger a cleaning operative. In this work, we design a methodology to train a *Convolutional Neural Network (CNN)* for identifying the level of obstruction in pipes. We gathered a database of videos to generate useful frames to fed into the model. Our resulting classifier obtains deployment ready performances. To validate the consistency of the approach and its industrial applicability, we integrate the *Layer-wise Relevance Propagation (LPR)* algorithm, which endows a further understanding of the neural network behavior. The proposed system provides higher speed, accuracy, and consistency in the sewer process examination.**

***Index Terms*—artificial intelligence, computer vision, pattern recognition, video recognition, deep learning, convolutional neural networks, explainability, sewers**

I. Introduction

In the US, there are roughly 1,200,000 kilometers of sewer lines [1]. That is more than three times the distance between the Earth and the Moon, considering only 4% of world population. The maintenance of such vast networks of pipes is thus a real challenge world-wide. As of now, the most common approach is to have operators executing sampling inspections, trying to find obstructions before they can cause severe failures that would require urgent and expensive actions.

The current approach is hardly scalable, as it is expensive and requires lots of human hours. Companies in charge of large wastewater networks face massive operational costs related to inspection and maintenance. The current environmental context brings added pressure to the topic since episodes of heavy rainfall are becoming more common because of climate change [2]. Within these episodes, obstructed wastewater networks may become the origin of sewer overflows and floods with an impact on urban environments and population.

To increase the quality and efficiency of sewer maintenance, the industry is now looking into recent technological advancements in fields such as image recognition and unmanned aerial vehicles. In this paper, we tackle one of the challenges necessary for new methods to be functional: the automatic identification of obstructions in sewer pipes from image data. For this purpose, we use real data from 6,590 inspection videos (samples shown in Fig. 1), recorded and evaluated by operators.

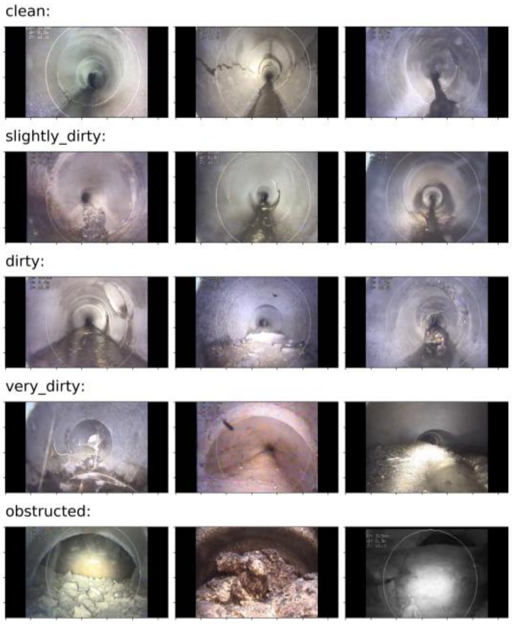


Figure 1. Sample frames from the videos database.

Manuscript received June 21, 2021; revised July 28, 2021; accepted August 5, 2021.

doi: 10.18178/ijscer.10.4.135-143

We post-process the videos to dissect and simplify the problem. With this data we devise, train, and evaluate a convolutional neural network (CNN) for predicting the level of obstruction of a sewer segment. The performance obtained in this work makes this technology suitable to the industrial challenge, increasing efficiency and enabling more extensive maintenance. In this case, this is already in progress through CETaqua, industrial partner of this project, and part of the SUEZ group.

II. Current Sewer Maintence

Regular sewer inspections are made for the operation and maintenance of the network. During inspections, the inside of sewers is recorded using a camera attached to a pole. Each video is carefully reviewed by an operator, who must fill a report and deliver it to the inspection site or to the central offices. The report must include the level of obstruction of the sewer, categorized into five classes: clean; slightly dirty; dirty; very dirty; and obstructed. Cleaning operations prioritize their interventions based on these reports.

Reviewing videos requires a significant amount of time from operators. This task is a major barrier for productivity because of its duration and repetitive nature; if the same operator dedicates too much time to this task, their performance will be affected. To avoid that, in practice, many different operators end up reviewing the same videos. While this is desirable for several reasons, it entails a significant variance in the evaluation criteria. Meanwhile, a reliable and consistent evaluation is critical for the efficient planning of maintenance.

Our goal is to define and implement a system to automatically assess the obstruction on sewers from videos. This system must provide a status on the volume of sedimentation to justify the cleaning needs. The deployment of this system in production will enable a more productive use of human resources and will provide a unified model for guiding cleaning operations.

III. State of the Art/Related Work

The use of computer vision techniques in civil engineering applications has grown exponentially, as visual inspections are necessary to maintain the safety and functionality of basic infrastructures. To mitigate the costs derived from the manual interpretation of images or videos, diverse studies explore the use of computer vision techniques. Methods like feature extraction, edge detection, image segmentation, and object recognition have been considered to assess the condition of bridges, asphalt pavement, tunnels, underground concrete pipes, etc., [3 4 5 6]. Moreover, noise reduction [7], and reconstruction and 3D visualization [8 9 10 11] were also used. In the most similar scenario to the one tackled in this paper, the automatic detection of cracks in sewers has been explored through image processing and segmentation methods [12 13].

Most of these related works are mainly focused on a single task, to detect cracks. Segmentation and classifications of pipe cracks, holes, laterals, joints, and collapse surfaces are explored through mathematical morphology techniques [14]. A most recent study uses these morphological operations and other pre-processing techniques, like edge detection and binarization, to identify the sewer defects by recognizing text displayed on the sewer video recording [15]. Even though computer vision techniques have provided a significant improvement in the analysis of civil infrastructure, there are still several difficulties to overcome, such as the extensive pre-processing of the data that must be carried out, a high degree of expert knowledge in the design of complex features extractors, the treatment of noisy and low-quality data, among others.

In this regard, CNNs require little image pre-processing, and more importantly, the feature extraction processes are learned automatically from the data through an optimization process. The performance of CNN models has been tested in several computer vision tasks, such as object detection or image classification. For instance, in the work of Y.-J. Cha *et al.* [16], an automated civil infrastructure damage system is presented, which is insensitive to the quality of the data and to camera specifications. Furthermore, CNNs use has demonstrated its efficiency in tunnel inspections [17], revealing how the deep learning approach outperforms conventional methods.

In the case of sewer inspections, the use of neural networks has been limited to defect detection. S.S. Kumar *et al.* proposes a convolutional neural network to identify root intrusions, deposits, and cracks in a set of sewer videos [18]. This database is transformed into a sequence of RGB images and fed them to the model. The training methodology is very straightforward, all images comprising a particular defect are feed to the CNN so that discriminative features can be learned. To enhance the performance of its model they used data augmentation, simulating a variety of conditions, and mitigating over-fitting. By doing so, the size of the dataset increases to millions of training samples for the model. However, and despite the good results, the model could not identify sub-classes, *e.g.*, fine roots from medium roots.

J. Cheng and M. Wang use a fast regional convolutional neural network (fast R-CNN) to detect different classes of sewer defects and to identify the coarse category to which they belong [19]. Their model is comprised of a set of images gathered from video sewer inspections which are fed to the model to generate both classification and bounding box regression of the defect. Despite is implemented data augmentation, the similarities in the geometry of the sewers and color gradients and intensity, penalized the model performance.

So far, there are no studies that discuss the automated classification of sewer obstruction level using CNNs. Previous works focus on more general faulty elements in the sewer structure, *e.g.*, roots, cracks, or deposits. However, due to the nature of the sewer system we work with (Barcelona area), it is crucial to assess if the sewer is free from obstacles, so that wastewater can flow through it ordinarily [20]. That being said, we can still use some of the insights found when tackling similar tasks.

IV. Sewer Data

CETaqua is a public-private research institution dedicated to the design of more sustainable water management services. Along the years, CETaqua has gathered a database of videos from 6,590 human-made inspections in the sewers of Barcelona area. Each video has an associated label, obtained from the operator's report which distribution is shown in Table I.

TABLE I. Original Video Distribution.

Label	# samples
clean	4720
slightly dirty	1146
dirty	235
very dirty	50
obstructed	98
Total:	6249

Sewer videos were obtained by different operators following a shared set of guidelines: the operator brings the camera down into the sewer and starts recording the pipe from a static position. After a few seconds, the operator zooms in to look further into the end of the sewer, followed by a zoom out to the starting position. At that point, the video ends. Videos are mostly shot at 360x640 resolution, and 10Fps (frames per second). Prototypical sample frames for every obstruction level are shown in Fig. 1. Since videos are recorded by different operators, there is a significant variance in their length. These range from ~18 to ~120 seconds. The distribution of video lengths is shown in Fig. 2. No video is excluded from this study because of its length.

Notice the difference in number of samples between the most frequent and the least frequent classes is of two orders of magnitude. Such large class imbalance may handicap the learning process of many machine learning algorithms, including CNNs. To tackle this issue, and following the advice of use case industrial experts, we merge the two classes with less samples per class, *very dirty* and *obstructed*, which are very close in meaning. We can do that without affecting the performance of the system because both classes imply the same industrial response once they are identified (i.e., prioritize cleaning of that sewer segment).

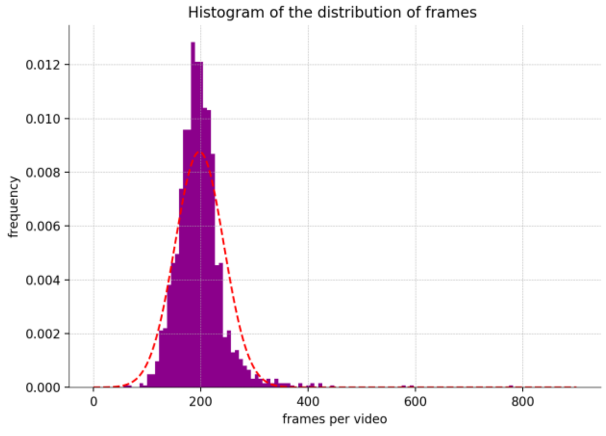


Figure. 2. Distribution of the video's length.

After the merging, the number of elements in the minority class has increased (to 148), but the uneven data distribution remains relevant, which could lead to a severe bias in the model performance. To avoid that we balance the distribution of the data by randomly *under-sampling* them to the minority class. Before data is fed into the model, we will still need to perform some pre-processing, to enable the learning process.

V. Dataset Engineering

The original task, as defined by the industrial requirements, is a video classification problem: Assign a given label to a given set of videos. However, we reduce this to an aggregated image classification problem to simplify it, as the inherent temporal aspect of videos is mostly irrelevant for our case. Working with images also increases the number of training samples, we can generate, as several frames from the same video become different (although not independent) training samples. With a larger training set we can improve the regularization and generalization of the CNN model.

Before transforming videos into images, we need to specify our dataset splits to avoid having images from the same video on both the training and test partitions. This would introduce a significant *bias* into the model, and significantly affect the relevance of our evaluation. After the *under-sampling* process, the distribution of the videos is shown in Table II. We have split the videos in two subsets: 70% for the training process and the remaining 30% for validation.

TABLE II. Videos Distribution per Dataset Split.

Label	Train samples	Validation samples
clean	103	45
slightly_dirty	103	45
dirty	104	44
very_dirty	104	44
Total:	414	178

A. *Frame Selection*

Of the full length of the video only a small portion of frames are usable for training. The zooming is digital on all cases, which means resolution is never increased, and some parts of the image are lost. For this reason, we gather the frames of the video where the camera is unzoomed. That is, from the beginning of the recording until the zooming in begins. To automatically locate this segment of interest we used the VidStab video stabilization algorithm from the OpenCV library [21]. This algorithm produces a smoothed trajectory of pixels using *key point* detectors like the examples in Fig. 3.

The top row examples of Fig. 3 are prototypical videos, where the zoom in and zoom out stages form a clear 'V'. Unfortunately, not all videos are like that. There is a significant variance and noise in the extracted trajectories, as shown in the bottom row of Fig. 3. Our analysis of trajectories shows that most videos have at least three seconds of image stability. Thus, we capture 30 consecutive frames for all videos. We do not extract a variable number of frames per video to avoid biasing the dataset.

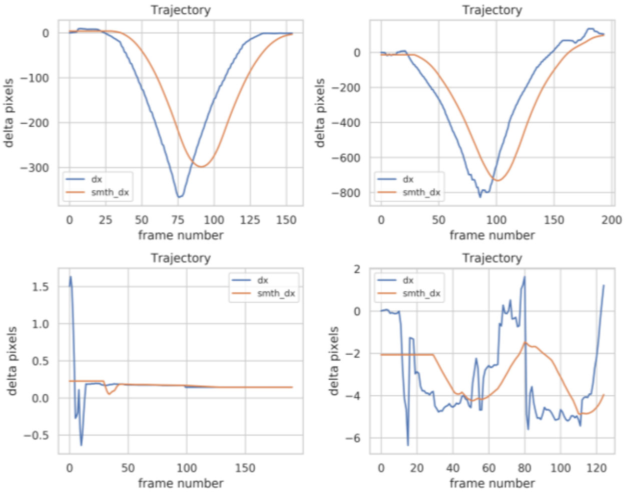


Figure 3. Samples of pixels trajectories. The blue line shows a change in pixel values. The red lines show a smoothed version of the same function. Notice the significant scale variations in the vertical axis.

We are using several frames per video which results in the distribution of Table III. Even though the number of images per class seems remarkable, this is deceptive. All images come from a hundred videos per class, which constrains the variance of our training set significantly. This also makes unproductive the use of generalization techniques like data augmentation, since there are already plenty of similar images with small variations in our dataset.

TABLE III. Images Distribution Per Dataset Split.

Label	Train samples	Validation samples
clean	3090	1350
slightly_dirty	3090	1350
dirty	3120	1320
very_dirty	3120	1320
Total:	12420	5340

B. *Input Pipeline*

Most frames have a resolution of 360x640. They also have a vertical border, as seen in Fig. 1. After removing it, images are at 360x480 resolution. For those few images that had a slightly higher resolution, we applied a central crop. During our experimentation, we noticed that models had the same performance if the 360x480 resolution was scaled down to 150x150. This is coherent with the task: since no specific object must be identified, fine-grained detail is unnecessary. For this reason, our final training dataset is composed of 150x150 images. Resizing the images also reduced the number of parameters needed and the training costs (*i.e.*, time, power and money).

VI. Models

NNs models are composed by a sequence of stacked layers which learn increasingly complex representations from the data. For image inputs, these representations are visual abstractions of shape, patterns, colors etc. which are used as building blocks for perception. In the context of our problem, where the goal is to identify the amount of obstruction, complex patterns are irrelevant for CNN. In other words, we do not care if the obstruction is caused by a bicycle or by a pile of cement.

What is essential to learn for the CNN is what a clean pipe looks like, and how different alterations to that normality correspond to different levels of obstruction. Clearly, spatial information is essential for the task, as sediment may be distributed along the channel, or it may form an obstruction at the bottom of the sewer. A sense of depth is also desirable, to assess obstructions proportions (and thus size) correctly. While we will not enforce these priors into the CNN, we will take them into account in our architectural designs, and we will validate them in our later interpretability study.

A. *Transfer Learning*

Fitting the many parameters found in deep CNNs to solve a task on high-dimensional inputs (*i.e.*, images) requires many data samples. To mitigate this need, one can use transfer learning: Initializing the parameters from a state optimized for a different problem, instead of initializing from a random state. Transfer learning is based on the assumption that most image challenges share a given set of visual properties which can be reused, instead of re-learnt. This is particularly true for low-level descriptors (*e.g.*, lines, angles etc.). Nevertheless, the transferability of features depends on the similarity between tasks. And in the variety and size of the data for which the pre-trained model was optimized [22]. For this reason, the most popular source models for transfer learning are those containing a wide variety of patterns (e.g., VGG16 [23]) for a wide variety of goals (*e.g.*, ImageNet [24]).

Considering the limited number of samples available in our task, we considered transfer learning as a potentially useful approach. We explore this hypothesis by using a VGG16 architecture trained on the ImageNet dataset. To adjust the VGG16 model to our needs, we start by removing the parameters of the original classifier (*i.e.*, the two fully-connected layers), since these are too optimized for the original problem and adapting the output of the network to fit our task. With this setting in place, we can now train the network through fine-tuning.

When fine-tuning, one must decide which layers to freeze (*i.e.*, fixing the weights), which to re-train (*i.e.*, fine-tuning the weights) and which to replace (*i.e.*, randomly initialized) or delete. The more layers we freeze, the more similar both tasks should be. Unfortunately, our case is a unique one, even when compared with a broad classification task like ImageNet. In our experiments, we gradually tried freezing a variable number of convolutional layers bottom-up. Significantly, none of these experiments was successful. In all experiments, the model either overfitted to the data or failed to learn meaningful representations. We hypothesize that the particularity of our problem makes it hard to reusing patterns learned on general-purpose datasets. Indeed, there is little in common between discriminating dog breeds and computing the level of obstruction of a sewer. On the other hand, the large number of parameters in networks trained for large tasks like ImageNet is inadequate for our small problem.

B. *Architecture Proposed*

Since transfer learning was unsuccessful, we defined an architecture design top to bottom for our problem. We started from a shallow architecture and increased its size until underfitting was no longer an issue. At that point, we optimized the hyper-parameters to get the best model. The Table IV shows the final CNN design. Notice the relatively small size of the architecture.

Increasing the number of filters and the kernel size provided no improvement, mostly because the variety of patterns to learn is small: the model does not have to recognize all possible objects and shapes that may obstruct the sewer. It must limit itself to learn what a clean sewer looks like, and what obstructions represent visually in that regard. Coherently, our experiments show that a CNN with only three convolutional layers and a fully-connected layer yields the best results. In our experiments, we used the *cross-entropy* loss function, ADAM optimizer with a 1e^{-6} learning rate, and 0.5 dropout value.

TABLE IV. CNN ARCHITECTURE PROPOSED.

Layer (type)	Output Shape	# Parameters
conv1 (Conv2D)	(150, 150, 32)	896
pool1 (MaxPooling2D)	(75, 75, 32)	0
conv2 (Conv2D)	(75, 75, 32)	9248
pool2 (MaxPooling2D)	(38, 38, 32)	0
conv3 (Conv2D)	(38, 38, 64)	18496
pool3 (MaxPooling2D)	(19, 19, 64)	0
flatten (Flatten)	(23104)	0
fc1 (Dense)	(1024)	23659520
dropout1 (Dropout)	(1024)	0
logits (Dense)	(4)	4100
Total params: 23,692,260		
Trainable params: 23,692,260		

VII. EVALUATION AND RESULTS

To evaluate the performance of the trained model, we got the confusion matrix. So, we can understand the frequency and severity of the mistakes made by the model. As shown in Fig. 4, 53.7% of images are classified in the correct class. 34.7% of images are classified in a neighboring class (*e.g.*, slightly dirty as dirty). The fact that mistakes are centered around the diagonal indicates that the model is properly learning the nature of the problem.

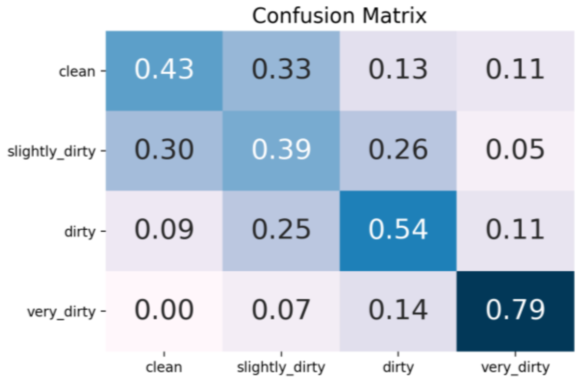


Figure 4. Normalized image-wise confusion matrix for validation set

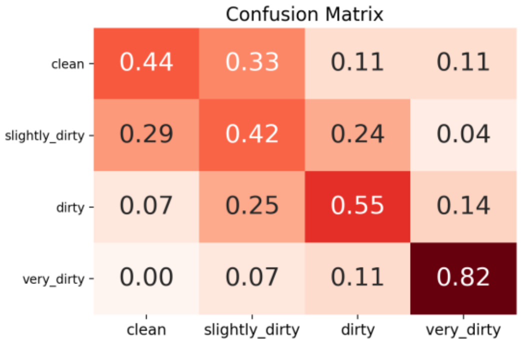


Figure 5. Normalized video-wise confusion matrix on the validation set. From the voting of classified images.

It is also worth noticing how the most relevant classes for the industrial application (the dirty and the filthy ones) are the ones classified with the highest accuracy. The last metric was computed image-wise, in the context of an image classification task. However, our ultimate purpose is to provide a video classification tool.

Based on the CNN image predictions, we generate a video classifier using a voting strategy, where each image from a video contributes with one vote towards the classification of the video itself. The confusion matrix of Fig. 5 shows the video-wise classification results. In this case, 55.7% of images are classified in the correct class, 2% more than the image classifier. The images classified in a neighboring class decrease from 0.7%, to 34.0%. The performance of this model fits the requirements of the industrial task.

Beyond the numeric analysis of the classifier outcome, we also explore in Fig. 6 some representative examples of failed predictions.

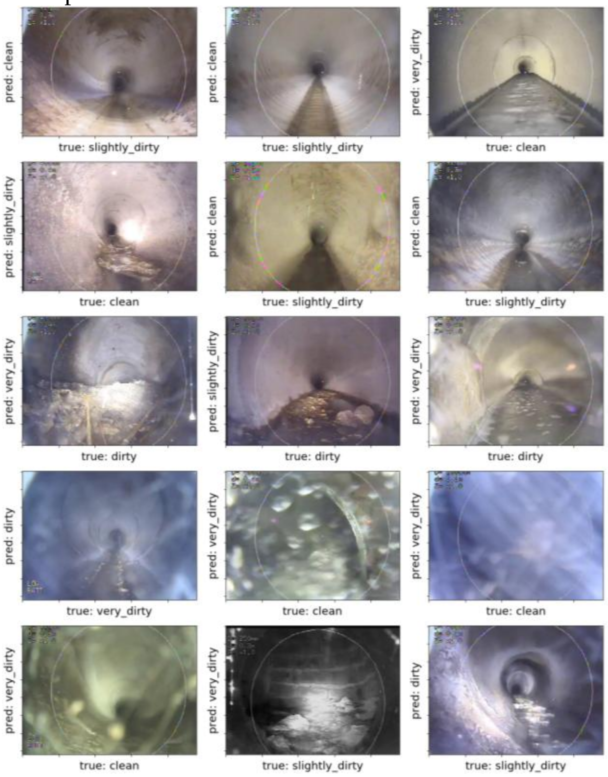


Figure 6. Samples of miss-predictions. true indicates ground truth. pred indicates model prediction.

The first two rows contain examples of videos where the labeling seems to be erroneous, which we attribute to human error. These samples could be re-labeled to improve the training dataset quality and the model performance.

The third row shows examples where the labeling criteria seems to be inconsistent because of having multiple operators labeling videos. Although the model predictions in these cases count as miss-classifications, its criteria seem adequately coherent. Another complicating factor we found in our analysis is rain like in the examples of the fourth row of Figure 6. Rain introduces lots of noise in the images, which handicaps perception and model prediction. Finally, the last row shows cases where the perspective of the camera is not normative (*i.e.*, centered in the pipe and looking towards the end of it). These variations confuse the model. To bypass this limitation more training data is needed.

A. *Interpretability*

So far, we have gathered evidence that the model is learning properly. Nevertheless, trusting the predictions of a black-box is never ideal. Explainability of the model is crucial for industrial risk assessment and regulation compliance. Thus, we take one more step into the validation of the model by looking at the visual patterns learned and used by the model to classify the data. This will provide interpretability to our system.

We integrate to our trained model the Layer-wise Relevance Propagation (LRP) algorithm to explore its decision-making process [25]. This algorithm tries to identify which features of the image input have the highest relevance for the prediction. Relevance is backpropagated from the output layer, assigning scores to the application of features, layer by layer until reaching the input. Each layer stores an equal amount of relevance, which is variably distributed among its features. The relevance of pixels in the input can be visualized through heatmaps as can be seen in the samples of the Figure 7.

For visibility reasons, LRP values are not normalized among all plots (*i.e.*, the same color on different LRP may indicate different relevance). The reference value for each LRP plot is shown above it (score relevance), and it depends on the confidence of that prediction. If all colors were normalized, colors from predictions with lower confidence would be barely visible. For this reason, plots of low probability predictions should not be over-interpreted.

Let us first consider what evidence is used to predict obstructions. As shown in the second column of Figure 7, the main evidence used for justifying a high level of obstruction (*i.e., dirty,* or *very dirty*) is located at the ground along the pipe. This seems adequate since this center canal will be naturally occupied by most obstructions. The LRP plots also indicate that changes in illumination are taken as evidence of obstruction (*e.g.*, third row, third column). Coherently, in a clean pipe light is smoothly distributed, while obstructed pipes contain segments of extreme illumination contrast.

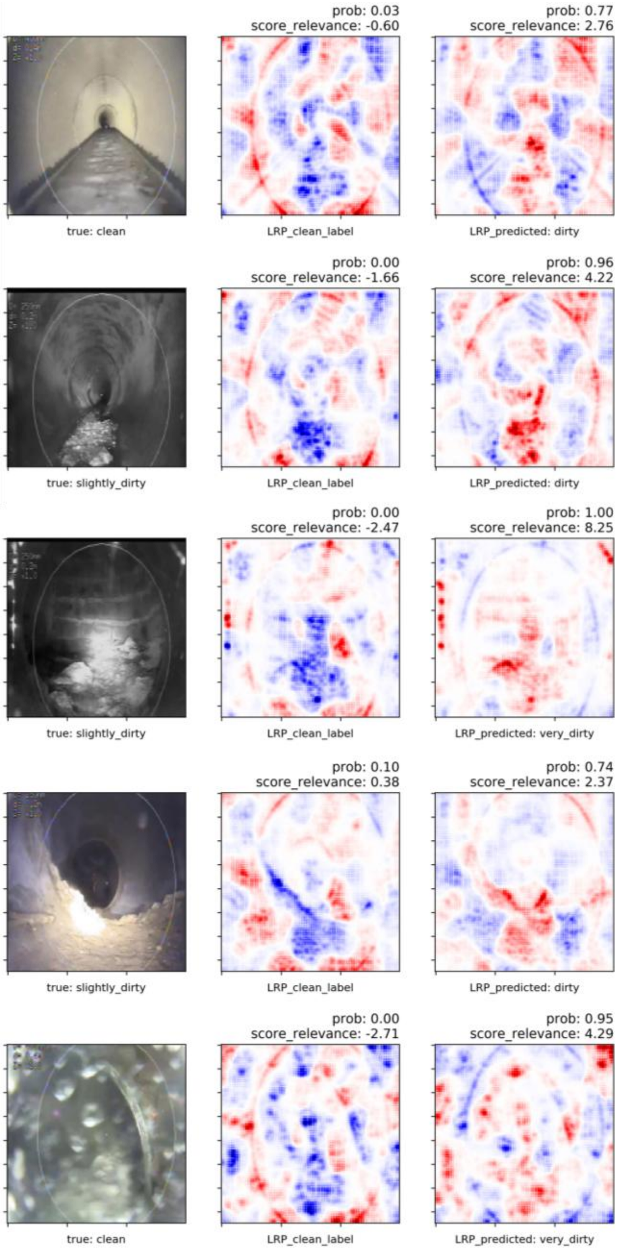


Figure 7. Each row shows an independent example. First column contains the original image. Second and third columns show the LRP of clean and dirty labels, together with the confidence of the prediction (score relevance). Red pixels indicate evidence in favor of prediction, blue pixels indicate evidence against it.

The use of both the ground path and illumination contrast as features for prediction explains the difficulties of the model for predicting images where there is either rain or changes in perspective. As shown in the bottom three rows of Fig. 7, the model still focuses on these features, even though in these cases such features characterize noise instead of obstructions.

VIII. Industrial Deployment

In this section, we outline the components for the implementation of the automated system for the evaluation of sewer conditions in the real environment. We design it so that the system keeps learning once deployed. The two main system components: a labeler API and a training pipeline.

The labeler API can be integrated into the maintenance department. It provides both automatic classification of videos, as well as a labeling interface for humans. Once a new video inspection is uploaded, the API is automatically requested for a classification. This will be done on a random number of frames from the static part of the video. The result, both classification and confidence, is processed by a rule-based system which determines what to do with the video. If the classification is *dirty* or *very dirty* and the confidence is high, send it to the cleaning team with urgency. If the confidence is low, send it to the queue for human labeling. If classification is *slightly dirty* or *clean* with high confidence, send it to the queue with low priority. All video labeled by humans through the interface are automatically used in the training pipeline.

The training pipeline is defined in a continuous integration server. When new videos arrive, these are fed into an object storage server. The storage server can trigger a series of jobs, after a minimum number of new samples are received. These jobs execute the following pipeline: 1. Dataset balancing and split; 2. Video stabilization and frame selection; 3. Frame resizing; 4. Model training; and 5. Model evaluation. The result of this pipeline is a model in *TensorFlow* along with a PDF document containing a sample of automatically labeled frames that are to be reviewed by an expert. If the results are good, the model is automatically deployed to the production API server, replacing the previous version. Every pipeline that generated every version of the model is stored along with the data used in it.

The system is designed for low-degree maintenance and for re-usability. The same pipeline could be potentially applied to any sewer system that shares strong similarities --*both structural and sedimental*-- with the one we have worked with. If differences were significant the CNN model architecture should be reassessed. It is therefore our assumption that this solution could be deployed internationally to any sewer management that uses video sampling inspections.

IX. Conclusions

The proper operation and the efficient and scalable identification of obstructions in sewer infrastructure is critical for current societies. In this context, operators are under pressure to record, evaluate, and perform inspections daily. In this work, we seek to alleviate this stressful task through a CNN model trained to identify the level of obstruction of a sewer.

We reduced the problem to an image classification one, as this is a more scalable and constrained approach. A pixel motion analysis allows us to measure the degree of noise in the dataset (which is high), and to define a unified frame extraction policy. Given a significant imbalance among target classes, we merge two similar classes and to down-sample the rest. In this setting we perform our experiments. Due to the limited data availability, we use transfer learning which failed, most likely, due to the dissimilarities between tasks. It remains to be seen if would be feasible a more flexible transfer learning mechanisms, like feature extraction where it is not needed to re-train the CNN [26].

In our experiments the best results are obtained by a rather small and shallow architecture, consistently with the nature of the task: There is no need to learn any specific pattern, just an overall sense of space and obstruction. The evaluation indicates this model learns to solve the task satisfactorily and illustrates the main reasons behind the failed predictions. Most frequently, inconsistent human labeling, variations in perspective and environmental noise like rain. We explore the behavior of the model by looking at the relevance of input pixels for output classification. This allows us to validate the visual features used by the model to make predictions. We notice how the center canal of the sewer is essential for the assessment of obstructions, how the visibility of circles around the pipe speaks for cleanness, and how changes in illumination and perspective can complicate the resolution of the problem.

Two more complicating factors were identified in the data during the development work. First, human mistakes when labeling videos. These are unfortunately frequent and bias the model performance. Second, the variability in labeling criteria. This is one of the motivating factors of this work, as an unstable policy reduces the quality and efficiency of maintenance interventions. Finally, beyond the visual model, we propose an integral system design to deploy all desirable functionalities.

Conflict of Interest

The authors declare no conflict of interest

Author Contributions

Mario A Gutierrez-Mondragon performed the design of the model, analyzed the data, planned, and carried out the experimentation and computations. He also conducted the analysis and interpretation of the results and took the lead in writing the manuscript. Dario García-Gasulla and Sergio Alvarez-Napagao were involved in planning and supervised the work and contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript. Jaume Brossa-Ordoñez and Rafael Gimenez-Esteban helped to supervise the project and validate the results according to the industrial requirements. All authors had approved the final version.

Acknowledgment

This work is partially supported by the Consejo Nacional de Ciencia y Tecnologia (CONACYT), Estudiante No. CVU: 630716, by the RIS3CAT Utilities 4.0 SENIX project (COMRDI16-1-0055), cofounded by the European Regional Development Fund (FEDER) under the FEDER Catalonia Operative Programme 2014-2020. It is also partially supported by the Spanish Government through Programa Severo Ochoa (SEV-2015-0493), by the Spanish Ministry of Science and

Technology through TIN2015-65316-P project, and by the Generalitat de Catalunya (contracts 2017-SGR-1414).

REFERENCES

[1] R Sterling, J Simicevic, E Allouche, W Condit, and L Wang, "State of technology for rehabilitation of wastewater collection systems," *Rep.EPA/600/R-10/078, US Environmental Protection Agency, Washington, DC. (Mar. 25, 2012)*, (2010).

[2] M. G. Donat, An. L. Lowry, L. V. Alexander, P. AO'Gorman, and N. Maher, "Addendum: More extreme precipita-tion in the world's dry and wet regions," *Nature Climate Change*, vol. 7, no. 2, pp. 154, 2017.

[3] I. Abdel-Qader, O. Abudayyeh, and M. E. Kelly, "Analysis of edge-detection techniques for crack identification in bridges," *Journal of Computing in Civil Engineering*, vol. 17, no. 4, pp. 255–263, 2003.

[4] J. Mashford, D. Marlow, D. Tran, and R. May, "Prediction of sewer condition grade using support vector machines," *Journal of Computing in Civil Engineering*, vol. 4, no. 25, pp. 283–290, 2011.

[5] H Zakeri, Fereidoon Moghadas Nejad, and Ahmad Fahimifar, "Image based techniques for crack detection, classification and quantification in asphalt pavement: a review," *Archives of Computational Methods in Engineering*, vol. 24, no. 4, pp. 935–977, 2017.

[6] P. Rose, B. Aaron, D. E. Tamir, L. Lu, J. Hu, and H. C. Shi, "Supervised computer-vision-based sensing of concrete bridges for crack-detection and assessment," *Transportation Research Board 93rd Annual Meeting*, Washington DC, 2014.

[7] T. Yamaguchi, S. Nakamura, R. Saegusa, and S. Hashimoto, "*Image-based crack detection for real concrete surfaces*," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 3, no. 1, pp. 128–135, 2008.

[8] S. Esquivel, R. Koch, and H. Rehse, "Reconstruction of sewer shaft profiles from fisheye-lens camera images," in *Proc. Joint Pattern Recognition Symposium*, pp. 332–341. Springer, 2009.

[9] D. Lattanzi and G. R. Miller, "3D scene reconstruction for robotic bridge inspection," *Journal of Infrastructure Systems*, vol. 21, no. 2, 2014.

[10] P. Huynh, R. Ross, A. Martchenko, and J. Devlin, "3D anomaly inspection system for sewer pipes using stereo vision and novel image processing," in *Proc. 2016 IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*, pp. 988–993. IEEE, 2016.

[11] C. Belles and F. Pla, "A Kinect-based system for 3D re-construction of sewer manholes," *Computer-Aided Civil and Infrastructure Engineering*, vol. 30, no. 11, pp. 906–917, 2015.

[12] M. R. Halfawy and J. Hengmeechai, "Efficient algorithm for crack detection in sewer images from closed-circuit television inspections," *Journal of Infrastructure Systems*, vol. 20, no. 2, 04013014, 2013.

[13] S. Iyer and S. K. Sinha, "Segmentation of pipe images for crack detection in buried sewers," *Computer-Aided Civil and Infrastructure Engineering*, vol. 21, no. 6, pp. 395–410, 2006.

[14] S. K. Sinha and P. W. Fieguth, "Morphological segmentation and classification of underground pipe images," *Machine Vision and Applications*, vol. 17, no. 1, pp. 21, 2006.

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

[16] Y. J. Cha, W. Choi, and O. B¨uy¨uk¨ozt¨urk, "Deep learning-based crack damage detection using convolutional neural networks," *Computer-Aided Civil and Infrastructure Engineering*, vol. 32, no. 5, pp. 361–378, 2017.

[17] R. Montero, J. G. Victores, S. Martinez, A. Jardon, and C. Balaguer, "Past, present and future of robotic tunnel inspection," *Automation in Construction*, vol. 59, pp. 99–112, 2015.

[18] S. S. Kumar, D. M. Abraham, M. R. Jahanshahi, TomIseley, and J. Starr, "Automated defect classification in sewer closed circuit television inspections using deep convolutional neural networks," *Automation in Construction*, vol. 91, pp. 273–283, 2018.

[19] J. CP Cheng and M. Z. Wang, "Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques," *Automation in Construction*, vol. 95, pp. 155–171, (2018).

[20] Francois Chataigner, Pedro Cavestany, Marcel Soler, Carlos Rizzo, Jesus-Pablo Gonzalez, Carles Bosch, Jaume Gibert, Antonio Torrente, Raul Gomez, and Daniel Serrano, "Arsi: An aerial robot for sewer inspection," *in Advances in Robotics Research: From Lab to Market*, pp. 249–274, Springer, (2020).

[21] G. Bradski, "The OpenCV Library", *Dr. Dobb's Journal of Software Tools*, (2000).

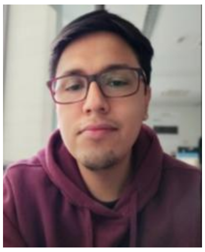
[22] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *Advances in Neural Information Processing Systems*, pp. 3320–3328, 2014.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprintarXiv:1409.1556*, 2014.

[24] H. Azizpour, A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "Factors of transferability for a generic convnet representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 9, pp. 1790–1802, 2015.

[25] S. Bach, A. Binder, G. Montavon, F. Klauschen, K. R. Muller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS One*, vol. 10, no. 7, e0130140, 2015.

[26] D. Garcia-Gasulla, A. Vilalta, F. Pares, E. Ayguade, J. Labarta, U. Cortes, and T. Suzumura, "An out-of-the-box full-network embedding for convolutional neural networks," in *Proc. 2018 IEEE International Conference on Big Knowledge (ICBK)*, pp.168–175. IEEE, 2018.



Mario A. Guti érrez is a PhD student in Artificial Intelligence at the Universitat Polytecnica de Catalunya (UPC), Barcelona Spain. The research area of the author is Machine Learning (ML) in general and Deep Learning (DL) particularly. He completed in 2016 a two-year master's degree program in Electrical Engineering in the Centro de Investigaci ón y Estudios Avanzados del Instituto Politecnico Nacional (Cinvestav I.P.N, Mexico City). He also worked during two terms (2016-2018) as an administrator of the Abacus-Cinvestav cluster in Mexico. MSc Mario A. Gutierrez has algorithmic and theoretical knowledge as well as computational experience in mastering different ML techniques, and the use of high-performance systems.

Dario Garcia-Gasulla received his PhD in Artificial Intelligence from Universitat Polytecnica de Catalunya in Barcelona, Spain in 2015. He is currently post-Doctoral fellow, leading research on large scale graph mining projects. These include cognitive architectures through deep learning and graph mining, network analysis for biomedical research, and scalability and optimization of parallel graph computing. He is also lecturer of the Deep Learning course in the Master in Artificial Intelligence (UPC-BarcelonaTECH).

Sergio Alvarez-Napagao received his PhD in Artificial Intelligence from Universitat Polytecnica de Catalunya in Barcelona, Spain in 2016.

Jaume Brossa-Ordo ñez received his MSc in Artificial Intelligence from IUBH Internationale Hochschule in Bad Honnef, Alemania. He is a Mathematician experienced in developing Computer Vision algorithms using Deep Learning techniques. Mr. Jaume is an enthusiastic professional to applying state-of-the-art Deep Learning algorithms to real industry challenges.

Rafael Gimenez-Esteban is a Computer Engineer from Universitat Polytecnica de Catalunya in Barcelona, Spain. He is a Big Data Engineer and Technical manager with solid vision and communication skills. He is also a data practitioner, Software Engineer and Senior Researcher. Manager of the 4.0 team at Cetaqua, a research group working to improve the efficiency and sustainability of the water cycle through state-of-the-art Artificial Intelligence.