Combining classifiers to detect faults in wastewater networks

Joshua Myrans, Zoran Kapelan and Richard Everson

ABSTRACT

This work presents a methodology for automatic detection of structural faults in sewers from CCTV footage, which has been improved by combining the outputs of different machine learning techniques. The predictions of support vector machine and random forest classifiers are combined using three distinct techniques: 'both', 'most likely' and 'stacking'. Each technique is tested on CCTV data taken from real surveys covering a range of pipes at locations in the south-west of the UK. The best tested technique, stacking, offers a 5% increase in accuracy for minimal impact in efficiency, proving useful for future development and implementation of the fault detection methodology.

Key words | classifier combination, fault detection, stacking

Joshua Myrans (corresponding author)
WISE Centre for Doctoral Training,
University of Exeter,
Harrison Building, North Park Road, Exeter,
EX4 4QF, Devon,
UK
E-mail: *jm494@exeter.ac.uk*

Zoran Kapelan
Centre for Water Systems,
University of Exeter,
Harrison Building, North Park Road, Exeter,
EX4 4QF, Devon,
UK

Richard Everson
Department of Computer Science,
University of Exeter,
Harrison Building, North Park Road, Exeter,
EX4 4QF, Devon,
UK

INTRODUCTION

Water companies in the UK must perform routine inspections to ensure the effective management and maintenance of their wastewater systems. Across the industry, it is standard to perform such inspections using CCTV cameras, particularly when pipes are too small for engineers to enter. These CCTV cameras can be remotely controlled by a technician, or pushed through the system using a series of rods. Both methods record detailed footage of the pipe's interior, used to identify faults within the system, recording the distance travelled along the pipe by measuring the length of tethering cable expended. It is common for the recorded footage to be analysed after collection, especially when pushing the camera through a system. This requires the entire length of footage to be watched by an engineer, who annotates the footage with identified faults. Alternatively, If the camera is being remotely controlled, a skilled technician may be able to annotate the footage as it is recorded. However, this often slows the survey as the camera is frequently stationary whilst a fault is annotated. In the UK annotation is performed according to the *Manual of Sewer Condition Classification* (WRc plc 2013), which is comparable to the industry standards across the world.

Most surveys, regardless of the collection method, are costly, requiring a team of technicians and expensive hardware. The surveying process is also a lengthy procedure; accessing the system, collecting the footage and annotating faults often takes days even for small surveys (<1 km of pipes). This can cost a company even more if they must close a section of road to access an asset, paying fees to local councils for the inconvenience. Given water companies' extensive wastewater networks (thousands of kilometres), CCTV surveys are in constant demand. In combination with their expensive and time-consuming nature, companies are always looking for ways to improve the process. This paper presents an automatic fault detection methodology, aiming to reduce the time required to analyse collected footage.

Previous studies into the automatic detection of sewer faults have yielded respectable results utilising an array of techniques. The most notable contributions have been made by Duran *et al.* (2003) and Sinha & Fieguth (2006); both utilised the power of artificial neural networks (ANNs) to identify faults within wastewater pipes. Duran *et al.* (2003) chose to augment cameras with a laser profiler, recording the internal dimensions of the pipe to generate a map of its surface. This map alone was passed to a trained ANN, which predicted the presence of a fault. The methodology was then extended to predict the category and severity of the fault using the same

doi: 10.2166/wst.2018.131

principle (Duran *et al.* 2007). This work was successfully demonstrated in laboratory-based experiments, fabricating faults in PVC pipes. On the other hand, Sinha & Fieguth (2006) chose to use standard CCTV footage in order to identify structural faults within pipes. The methodology used average light intensity to identify regions of interest, before extracting a series of key features, such as light intensity, shape and size. These measured features were combined with knowledge of the problem using fuzzy logic, before being again passed to a series of ANNs. Unlike Duran *et al.* (2007) this work was demonstrated on real-life CCTV footage, collected from flush-cleaned pipes. More recently, Halfawy & Hengmeechai (2014) examined regions of interest within images taken from CCTV surveys, extracting histogram of ordered gradient features for each area of interest, before using a support vector machine (SVM) to predict the presence of a fault. This produced good results for the tested categories of fault, although the methodology was not demonstrated on a collection of subtler faults, including cracks, deposits and surface damage. In addition, Halfawy & Hengmeechai (2014) imply the use of a detector for each type of fault, with the collection of detectors being applied to every area of interest within an image.

In contrast to previous work, this paper aims to demonstrate a single holistic methodology, capable of detecting all faults within wastewater networks. The work presented here continues to develop work previously published (Myrans *et al.* 2016a, 2016b), in which the success of random forests (RF) and SVMs were demonstrated on real CCTV footage of uncleaned pipes. As each classifier showed improved performance on different categories of fault, the methodology presented here looks to combine the predictions of the classifiers in an intelligent manner, so as to utilise the best aspects of each. Unlike other methodologies, this methodology is demonstrated on a broad variety of faults in both still frames and continuous CCTV footage, all of which have been taken directly from real life CCTV surveys of uncleaned pipes.

METHODOLOGY

The overall fault detection concept is described in more detail in Myrans *et al.* (2016a, 2016b) and it consists of five stages: 'Extraction', 'Pre-processing', 'Feature extraction', 'Classification' and 'Interpretation'.

Fault detection

The first stage 'Extraction' simply requires individual frames to be lifted from the raw CCTV footage. This is done at a rate relative to the speed of the camera; however, the methodology is currently efficient enough to analyse every frame of footage in near real time. The second stage 'Pre-processing' resizes frames to a predetermined resolution. Frames are also converted from colour to greyscale, as the extra information provided in a colour image is often unnecessary in identifying faults. In combination, these two steps significantly reduce the number of values required to describe the image, dramatically simplifying the detection problem without loss of accuracy. The third stage 'Feature Extraction' continues reducing the complexity of the problem, extracting a GIST feature descriptor (Oliva & Torralba 2001). A GIST feature descriptor aims to describe the contents of an image in a much lower dimensional space. It does so by splitting the image into regions, which are each convolved with a series of Gabor wavelets, capable of representing directional patterns. After summing the contents of each region, a final feature descriptor is produced, which, following the specifications of Oliva & Torralba (2016), is a 512-dimensional vector. The fourth stage 'Classification' uses a machine learning classifier, either an RF or SVM, to make a prediction of the probability of a frame containing a fault. Both classifiers must be trained before use; this should be performed using a collection of relevant labelled GIST feature descriptors. The RF is an ensemble classifier, using a collection of randomly 'grown' decision trees to predict the classification (faulty or normal) of a given GIST descriptor (Breiman 2001). On the other hand, an SVM maps the GIST descriptor to a higher dimensional space, attempting to split the data into two classes using a separating hyperplane (Cortes & Vapnik 1995). The fifth and final stage of the methodology, 'Interpretation', uses the classifier's prediction to determine the presence of a fault. If working on isolated frames this is a simple case of setting a threshold on the classifier's output. Alternatively, if working with a sequence of video a hidden Markov model (HMM) is implemented, relating information between all frames in a sequence to provide a better-informed classification of a frame's contents (Rabiner 1989).

Combining classifiers

As both RF and SVM classifiers have proven to be successful where the other has failed, we look to apply both classifiers during the 'Classification' stage of the fault detection methodology with the aim of improving the prediction accuracy. By applying both an RF and SVM we look to combine their outputs in an intelligent manner before making a final classification in the 'Interpretation' stage. This paper

demonstrates three such techniques, comparing their success in the case study later.

The first method demonstrated is the simplest, taking the 'most likely' prediction from the pair of classifiers. We define the 'most likely' output to be the probability closest to either classification (0 or 1, i.e. no presence or presence of a fault predicted). By choosing the 'most likely' prediction we aim to choose the better performing classifier for each type of frame, using the chosen output as the observed state in the HMM (stage 5). However, this method relies on outputs from different models, which are hard to correlate and so could lead to error in the final prediction.

The second method relies on the ability of HMMs to implement multiple observations; as such we use the output from 'both' the RF and SVM as observations in the model. Providing more relevant observations improves the accuracy of the state transition matrix, governing whether a frame's true state is classified as faulty or normal (Rabiner 1989); in turn this should improve the accuracy of a frame's classification during the 'Interpretation' step of the fault detection methodology.

The third and final method attempts to intelligently combine the classifier's predicted probabilities using the 'stacking' technique (Sill *et al.* 2009). This technique implements a second level 'stacking' classifier to optimally combine the predictions of multiple different models. By passing the predicted probabilities of the RF and SVM to a further stacking classifier, the original outputs are combined, creating a prediction probability, influenced by both base classifiers. An extension to this technique passes the original feature vector to the stacking classifier, in addition to the predictions of the base classifiers (Sill *et al.* 2009). This is called feature weighted stacking (FWS) and aims to provide the stacking classifier with as much information as possible to make an accurate prediction. In both versions of this technique the stacking classifier must be trained on labelled examples; these examples should be different to those used to train the base classifiers.

Case study

To demonstrate the three combination techniques, each was applied to sequences of continuous CCTV video. The footage used in these experiments was provided by the UK water company Wessex Water, taken from their recent routine sewer surveys. The CCTV data cover a variety of wastewater pipes with total length of over 5.5 km, diameters ranging 150–1,500 mm and made from materials including brick, vitrified clay and concrete. From the CCTV video (total length of over 10 hours), a dataset of roughly 1,500 frames was extracted and labelled according to the *Manual of Sewer Condition Classification* (WRc plc 2013). Approximately half the frames contained faults; most faults common to UK water networks were represented including cracks, intruding roots and obstructions among many others, the distributions of which are shown in Table 1.

In addition to the dataset of labelled images, a collection of video sequences was set aside to effectively test each method. These sequences totalled 30 minutes of footage, and covered a variety of common faults including cracks, root intrusions and displaced joints. Each of these sequences was again manually labelled according to the engineer's annotations, enabling the objective analysis of each method.

To effectively compare all three techniques the base classifiers will also be evaluated, reviewing the methodology's performance prior to combination. In addition, four key decision-making criteria are observed enabling a fair comparison of the techniques:

- Accuracy – the number of frames classified in agreement with the surveyor's annotations. The surveyor's annotations are regarded as absolute truth during these experiments. It should be noted that annotations can be inconsistent and are often prone to human error as shown by van der Steen *et al.* (2014).

Table 1 | A table demonstrating the distribution of fault labels in the Wessex Water dataset

Fault Type	Subtypes	Distribution (%)
Deposits	Attached, Settled	32.7
Joint	Displaced, Open	30.0
Crack	Longitudinal, Circumferential, Multiple, Spiral	14.5
Broken/ Collapsed	–	10.1
Hole	–	6.8
Obstruction	Intruding junctions, Masonry, Protrusions	6.6
Brickwork	Missing mortar, Displaced bricks, Missing Bricks	3.9
Roots	Fine, Tap, Mass	3.7
Infiltration	Running, Gushing	2.8
Surface	–	2.2
Deformation	–	0.4

- Ratio of false positive rate (FPR) to false negative rate (FNR) – the rate at which frames are incorrectly classified by the method. FPR refers to the number of frames classified as faulty, where no fault is present, whilst FNR refers to the number of frames classified as normal where a fault is present. These metrics are key to having confidence in the methodology, a high FPR wastes a technician's time as frames are unnecessarily flagged. On the other hand, a high FNR is potentially worse, as the methodology ignores faults which would otherwise have been identified by a technician.
- Area under the curve (AUC) – the area under the receiver operating characteristic (ROC) curve. This represents the overall performance of the methodology, where the ROC curve visualises the trade-off between FPR and FNR achieved by varying internal classification thresholds. When comparing two methods one clearly outperforms the other, if it has a higher AUC and its ROC curve completely dominates the second method.
- Speed – the additional computational time required by the technique relative to the case where classifiers are not combined. With the overall goal of speeding up sewer surveys, if this additional step improves accuracy at the cost of significantly slower performance this cost/benefit trade-off should be considered. All techniques will be run on the same desktop computer with a quad core processor (Intel Core2 Quad @2.66 GHz).

RESULTS AND DISCUSSION

To make a fair comparison between methods, the performance of both base classifiers has been documented, in addition to each combination technique. Implemented on the 30 minutes of unseen CCTV footage, the RF classifier achieved an accuracy of 73%, whilst the SVM achieved an accuracy of 70%. Examining the confusion matrices in Table 2, we can see the SVM achieves a lower FPR of 36%, compared with the RF's 39%. On the other hand, the RF has a lower FNR of 17%, compared with the SVM's FNR of 26%.

Table 2 | Confusion rate matrices for the base SVM and RF

Base SVM (0.70)		Truth: Fault	Truth: Normal	Base RF (0.73)		Truth: Fault	Truth: Normal
Predicted	Fault	0.74	0.36	Predicted	Fault	0.83	0.39
	Normal	0.26	0.64		Normal	0.17	0.61

It should also be noted that increasing the size of the training set to 1,500 images (the original set of 1,000 + the 500 used to train the stacking classifier) had little impact on the accuracy of both base methodologies. With the additional training the SVM classifier achieved the same accuracy of 70% and the RF accuracy increased to 74%. For this reason, the same base classifiers, trained on 1,000 images, were used for all combination techniques.

The first combination technique, choosing the most likely prediction, achieved an accuracy of 74%. A marginal improvement on the RF, it achieved the same FPR of 39%, whilst reducing the FNR to 16% (Table 3). Implementing this technique had a negligible impact on the processing speed, adding 0.14 seconds of processing time per hour of footage analysed.

The second technique, using predictions from both base classifiers, achieved an accuracy of 72%, performing slightly worse than the base RF classifier. The confusion matrix in Table 4 shows the averaging effect between the two base classifiers, achieving an FPR of 37% and an FNR of 20%. Again, implementing this technique had negligible impact on the processing speed, adding 0.04 seconds of processing time per hour of footage analysed.

The third technique was implemented in both forms, using stacking and FWS. The techniques used an SVM

Table 3 | Confusion rate matrix for the 'most likely' combination technique

Likely (0.74)		Truth: Fault	Truth: Normal
Predicted	Fault	0.84	0.39
	Normal	0.16	0.61

Table 4 | Confusion rate matrix for the 'both' combination method

Both (0.72)		Truth: Fault	Truth: Normal
Predicted	Fault	0.80	0.37
	Normal	0.20	0.63

classifier, separately trained on 500 additional labelled frames. These frames were taken from surveys not used in the base classifiers' training, nor in the 30-minute CCTV segment used for testing. Both methods demonstrated an improved accuracy of 78%, 5% better than the base RF classifier. The stacking technique achieved an FPR of 30% and an FNR of 15%, whilst the FWS achieved an FPR of 23% and FNR of 22% (Table 5). Implementing the stacking method, excluding training, added 0.5 seconds of processing time per hour of footage analysed, whilst FWS added 4.7 seconds of processing time per hour of footage analysed. Given that the process methodology currently runs in near real time, taking just over an hour to process 1 hour of footage, this extra processing could also be considered negligible.

The final criterion is demonstrated in Figure 1. It shows the base classifiers achieved an AUC of 0.77 and 0.83 for the SVM and RF respectively. Choosing the most likely prediction achieved an AUC of 0.82, whilst using both predictions achieved an AUC of 0.8. Improving on the base classifiers, the stacking technique achieved an AUC of 0.84, whilst FWS achieved the highest AUC of 0.85.

As an overall methodology all of the combination techniques have proven extremely viable. Upon manual inspection of each technique's shortcomings, subtle, less severe faults are the most commonly misidentified. For regular sewer inspections a TPR of 0.85 is more than acceptable, comparable to human error in current inspections (van der Steen *et al.* 2014). However, due to expenses, most sewer surveys are performed on high risk pipes (Mashford *et al.* 2010). In these cases, the TPR of all techniques could be improved (to >90%) at the cost of a higher FPR; this trade-off is best demonstrated by the ROC curves in Figure 1. This flexibility allows the methodology to be tuned on the fly, being cautious on high risk sewers to further reduce the chance of a fault being missed.

In this case study, stacking and FWS appear to perform the best, providing the largest improvement in three of the four decision-making criteria. Both FWS and stacking achieved the highest accuracy of 0.78, as well as the lowest FPR and FNR and best AUCs. The two considerations of these stacking techniques are the minor increase to processing time and the requirement for additional training data. However, in practice these additional costs should have negligible impact. The

Table 5 | Confusion rate matrix for the stacking and FWS method

		Truth				Truth	
Stacking (0.78)		**Fault**	**Normal**	**FWS (0.78)**		**Fault**	**Normal**
Predicted	Fault	0.85	0.30	Predicted	Fault	0.78	0.23
	Normal	0.15	0.70		Normal	0.22	0.77

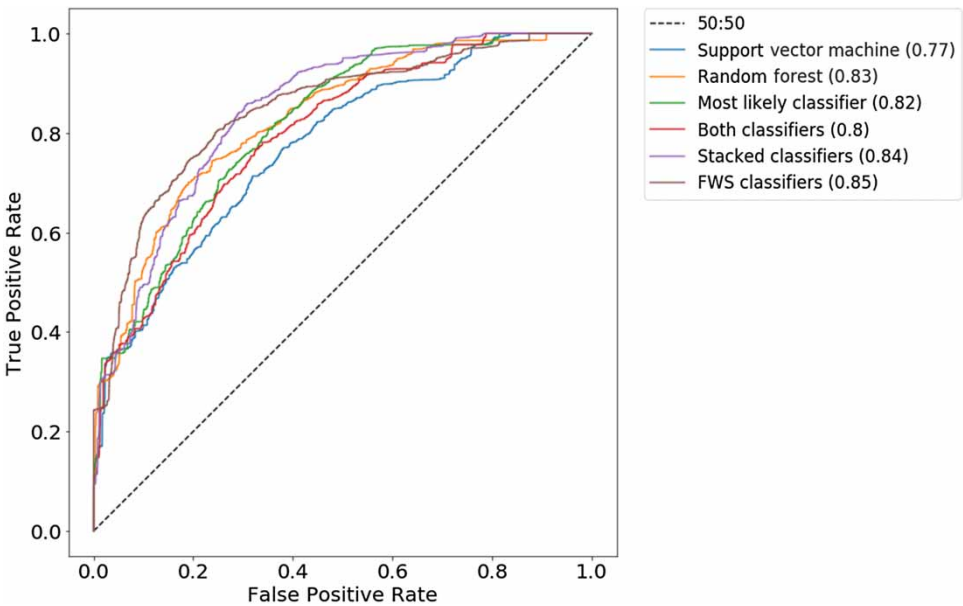

Figure 1 | ROC curves for each combination technique and base classifier. Note the 50:50 line represents guessing at a frame's contents and is drawn only for comparison purposes.

additional processing time still leaves the methodology capable of real-time predictions and most water companies who may use this technology will have an abundance of archived CCTV footage which could be used for training.

On the other hand, distinguishing between stacking and FWS is much harder, as neither clearly dominates the other. Both achieve the same accuracy, whilst stacking has a lower FNR and FWS has a lower FPR. Normal stacking may be slightly faster; however, FWS has a slightly higher AUC and dominates for more of the ROC plot. For these reasons, either could be implemented in future iterations of the overall fault detection methodology and one should be selected on a case-by-case basis.

CONCLUSION

Knowing that RF and SVMs perform well in distinct aspects of the fault detection methodology, this paper demonstrated multiple techniques for combining the outputs of both machine learning techniques. The techniques presented and tested included: the selection of the most likely prediction, using both predictions as observations in the 'Interpretation' stage of the fault detection methodology and stacking, including FWS. The techniques were compared using four key performance criteria: accuracy, FPR/FNR, speed and AUC. Performing the best, stacking and FWS increased the accuracy of the methodology by 5%, reducing both FPR and FNR whilst only marginally slowing the entire fault detection methodology. Continuing to compare all methodology's ROC curves and AUC, both stacking and FWS dominated the other techniques. Using both SVM and RF predictions appeared to average between the two classifiers, resulting in no net improvement. On the other hand, selecting the most likely classifier showed small improvements over the more successful RF classifier. Both stacking and FWS were recommended as improvements to the fault detection methodology, as both perform exceptionally and have only subtle differences.

Future work in this area will involve further extensive testing of the fault detection methodology, implementing the tool on larger and more complex datasets. Work will also begin on the extension of the fault classification methodology, starting to categorise faults by type, location and severity. Ultimately, this work aims to combine these technologies to create a user-friendly decision support tool, capable of assisting technicians in the field.

ACKNOWLEDGEMENTS

This work was supported by the Engineering and Physical Sciences Research Council in the UK via grant EP/L0116214/1 awarded for the Water Informatics, Science and Engineering (WISE) Centre for Doctoral Training, which is gratefully acknowledged. This work was also kindly supported by Wessex Water (Julian Britton) who provided the annotated CCTV footage and industrial insight, which is equally gratefully acknowledged.

REFERENCES

Breiman, L. 2001 Random forests. *Machine Learning* **45** (1), 5–32.

Cortes, C. & Vapnik, V. 1995 Support-vector networks. *Machine Learning* **20** (3), 273–297.

Duran, O., Althoefer, K. & Seneviratne, L. 2003 Pipe inspection using a laser-based transducer and automated analysis techniques. *IEEE/ASME Transactions on Mechatronics* **8** (3), 401–409.

Duran, O., Althoefer, K. & Seneviratne, L. D. 2007 Automated pipe defect detection and categorization using camera/laser-based profiler and artificial neural network. *Automation Science and Engineering IEEE Transactions on* **4** (1), 118–126.

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

Mashford, J., Marlow, D., Tran, D. & May, R. 2010 Prediction of sewer condition grade using support vector machines. *Journal of Computing in Civil Engineering* **25** (4), 283–290.

Myrans, J., Kapelan, Z. & Everson, R. 2016a Automated detection of faults in wastewater pipes from CCTV footage by using random forests. *Procedia Engineering* **154**, 36–41.

Myrans, J., Kapelan, Z., Everson, R. & Britton, J. 2016b Using Support Vector Machines to identify faults in sewer pipes from CCTV surveys. In: *CCWI 2016, Electronic Proceedings*.

Oliva, A. & Torralba, A. 2001 Modelling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision* **42** (3), 145–175.

Rabiner, L. R. 1989 A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77** (2), 257–286.

Sill, J., Takács, G., Mackey, L. & Lin, D. 2009 Feature-weighted linear stacking. arXiv preprint arXiv:0911.0460. Cornell University, Ithaca, NY, USA.

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

van der Steen, A. J., Dirksen, J. & Clemens, F. H. 2014 Visual sewer inspection: detail of coding system versus data quality? *Structure and Infrastructure Engineering* **10** (11), 1385–1393.

WRc plc 2013 *Manual of Sewer Condition Classification*, 5th revised edn. WRc Publications, Swindon, UK.

First received 27 October 2017; accepted in revised form 10 March 2018. Available online 20 March 2018