

ANALYSIS AND CLASSIFICATION OF INTERNAL PIPELINE IMAGES

Deirdre O'Brien, Maya Gupta, Robert M. Gray

Jon Kristian Hagene

Stanford University, Dept. of Electrical Engineering
Stanford, CA 94305
{dbobrien, guptama, rmgray}@stanford.edu

Norsk Elektro Optikk A/S
Lørenskog, Norway
jonkr@neo.no

ABSTRACT

Recently developed optical inspection tools provide images from the inside of natural gas pipelines to monitor pipeline integrity. The vast amounts of data generated prohibits human inspection of the resulting images. We designed an image processing and classification method to identify abnormal events. Non-overlapping image blocks are classified into twelve categories: normal, black line, grinder marks, magnetic flux leakage inspector marks, single dots, small black corrosion dots, osmosis blisters, corrosion dots, longitudinal weld, field joint, cavity at a weld and longitudinal weld too close to field joints. Results compare different types of statistical classifiers. Features extracted from the pipeline image are designed to mimic the features humans use to identify the different classes. Difficulties include the large number of classes, the uneven costs associated with different errors, and training on a limited amount of expert classified data. Classification results show this to be a useful tool for pipeline monitoring.

1. INTRODUCTION

Natural gas pipelines in the North Sea and on land in continental Europe are aging and identification of signs of corrosion and decay is important. Monitoring the pipeline integrity can avert costly leaks and speed repairs in the event of an accident. Currently, the primary technology for inspecting these pipelines is magnetic flux detectors [1]. Such inspections are expensive, difficult, and at times inaccurate. A new optical inspection technology has been developed by Norsk Elektro Optikk (NEO) which shuttles a laser camera through the pipelines storing images of the inner walls of the entire pipeline. It is inefficient to manually analyze the resulting kilometers of data. A goal is to develop an automatic system for analyzing the images and identifying anomalous events, thus providing information on the overall pipeline integrity and areas of the pipeline in need of repair.

This work was partially supported by Norsk Elektro Optikk, by NSF Grant No. CCR-0073050 and by an NSF Graduate Fellowship.

The original dataset is kilometers of JPEG compressed image data. These are broken into 96×128 pixel images, each representing sections of pipeline approximately $96 \text{ mm} \times 128 \text{ mm}$. Examples of the images are shown in Figure 1, the background vertical stripes in the images are due primarily to variations on the laser intensity across the line imaged by the camera.

A hand-labelled database of 228 images was created representing twelve classes of pipeline events reflective of pipeline integrity. As shown in Table 1, the database contains an uneven spread of the twelve classes.

In Section 2, we explain the design of the 22 features used to represent each 96×128 image. We discuss the different classifiers evaluated in Section 3 and detail the experiments in Section 4. The results in Section 4 show that different classifiers have different advantages and can achieve good performance recognizing abnormal events.

Class Label	Description	Number of Images
A	Normal	43
B	Osmosis Blisters	20
C	Black Lines	14
D	Small Black Corrosion Dots	17
E	Longitudinal Welds	20
F	Weld Cavity	19
G	Welds Too Close	16
H	Field Joint	20
I	Grinder Marks	20
J	MFL Marks	13
K	Corrosion Blisters	11
L	Single Dots	15

Table 1. Frequencies of the twelve pipeline event classes.

2. FEATURES

Features based on raw pixel values, DCT coefficients, and wavelet coefficients for 8×8 blocks were explored early in the project, but did not capture enough information to

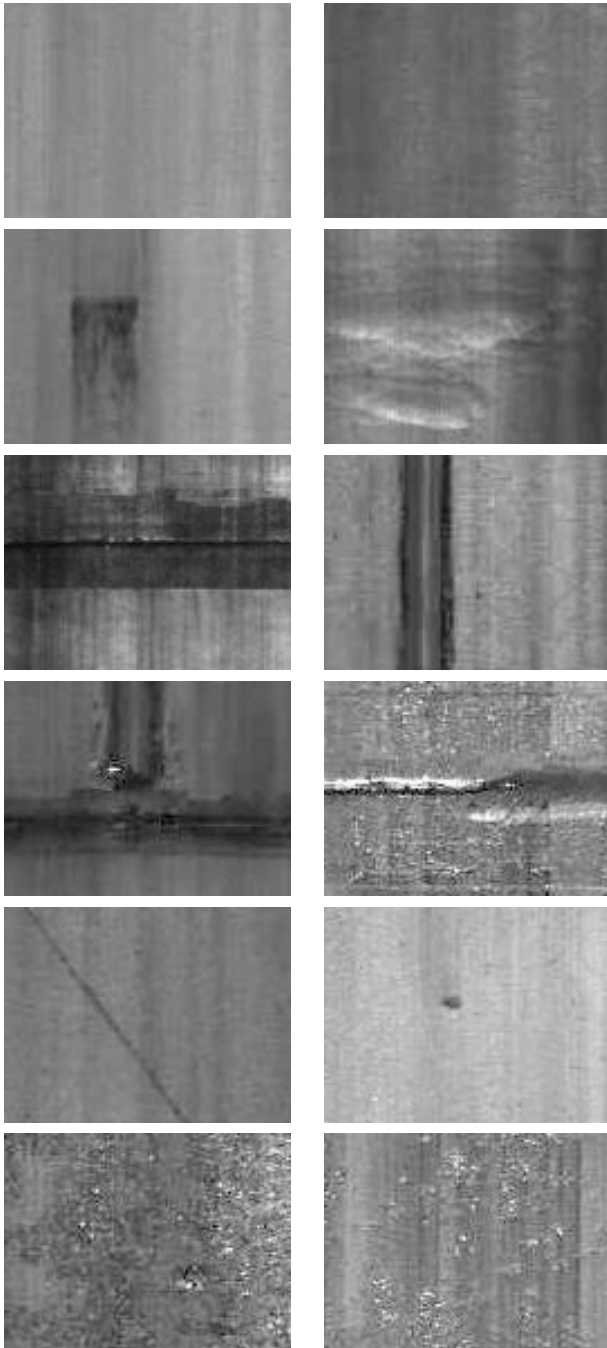


Fig. 1. Example 96×128 pipeline images. Left to right, top row: normal, normal. 2nd row from top: MFL mark, grinder mark. 3rd row from top: field joint, longitudinal weld. 4th row from top: welds too close, weld cavity. 5th row from top: black line, single dot. Bottom row: corrosion blisters, osmosis blisters.

achieve useful classification results. The events of interest generally occupied areas significantly larger than 8×8 blocks. Also, the artifacts due to laser variation and those resulting from the JPEG compression detracted from the information that could be extracted using this small blocksize.

Nonlinear, statistical, and morphological features were found to yield better differentiation between classes. Designing specific features for this particular application allowed prior knowledge of the artifact appearance to be incorporated into the features. This reduced the expected overfitting due to the small dataset.

The features used fall into two categories – those designed using the contrast feature of gray-level co-occurrence matrices (GLCM) [2] and those which we term human visual discriminant (HVD) features.

Six GLCM contrast features were designed. First, local contrast was measured as the squared differences of pixels vertically separated by 4 mm or 8 mm, averaged over $4 \times 4 \text{ mm}^2$ image blocks. Then, regions of high contrast were identified, such that the local contrast is greater than a constant multiple of the image variance. Six features were extracted which describe the total area and number of these connected high contrast areas as well as the size of the largest such area.

The 16 HVD features seek to mimic the features used in manually distinguishing between classes. The first feature measured the standard deviation of image pixels. Nine features were extracted which capture information about the size, shape and mean graylevel of relatively dark or bright areas. Dark and bright pixels were identified as those with graylevels of a number of standard deviations from the image mean.

Other features were designed to specifically differentiate certain classes. To recognize *blisters*, a feature calculates the number of 8×8 blocks within the image with a range larger than a given threshold. To identify horizontal artifacts such as *field joints*, the average difference between the graylevel averages of each side of a horizontal boundary that sweeps down the image was included. To help find *corrosion dots*, one feature counts how many 4×4 blocks have means that are significantly darker than the column mean. Other features based on vertical strips through the image measure the mean of the 4 mm wide column standard deviations and the standard deviation of the mean of 2 mm wide columns as well as the maximum value of the image mean minus the mean of a 2 mm wide column. These final features are used to identify vertical artifacts such as *longitudinal welds*.

Each of these features were normalized to have unit variance. Unit variance feature scaling may not be the optimum relative scaling for the features. It is likely that certain features are better discriminants while others are noisy or less informative. One way to generate improved feature scal-

ings is to use the feature importance metric incorporated in the CARTTM (classification and regression trees [3]) algorithm. CART seeks to build a tree that minimizes the misclassification error. The importance of a particular feature can be measured by how closely a tree built using only that feature matches the best tree found by CART. Scaling the univariate features by CART’s feature importance measures significantly improved the classification performance.

3. CLASSIFICATION ALGORITHMS COMPARED

Four classifiers were compared: multiple additive regression trees (MART) [4], linear discriminant analysis (LDA) [4], regularized quadratic discriminant analysis (QDA) [4], and linear interpolation with maximum entropy (LIME) [5].

MART is a gradient boosted version of a classification tree.¹ MART estimates the probability that a test sample X belongs to class s , $\hat{P}_s(X)$. Given a cost matrix C , where $C(r, s)$ is the cost of assigning a sample to class s when it belongs to class r , the assigned class is

$$\arg \min_{\hat{y}} \sum_s \hat{P}_s(X) C(\hat{y}, s). \quad (1)$$

MART parameters were set at values suitable for this small dataset.

LDA fits a Gaussian with the same covariance to each class. QDA calculates the covariance independently for each class. Regularized QDA uses a weighted average of the LDA and QDA covariances for each class, providing a more general model than LDA while being more robust than QDA on the small dataset. Using this model the probability estimates $\hat{P}_s(X)$ were calculated and X assigned using equation (1).

Nonparametric methods, such as k-NN and kernel methods [4], can be severely biased by uneven distributions of training data. Linear interpolation with maximum entropy (LIME) [5] is a new nonparametric method that avoids these problems by weighting training samples based on their spatial relationships, using the equations of linear interpolation. LIME has two parameters to train, the cardinality of the neighborhood set k , and the trade-off λ between the linear interpolation equations and maximizing entropy.

LIME classifies a test feature vector X in three steps:

Step 1) Find the nearest k neighbors, and let (X_j, Y_j) denote the j th nearest neighbor.

Step 2) Calculate a probability distribution w such that $\sum_{j=1}^k w_j = 1$ and $w_j \geq 0$ for all $j = 1$ to k , and such that the weights solve

$$\arg \min_w \left[\left\| \sum_{j=1}^k w_j X_j - X \right\|_2 - \lambda H(w) \right]$$

¹MART was implemented using code available at <http://www-stat.stanford.edu/~jhf/>

where λ is a chosen parameter, $\|\cdot\|_2$ is the L_2 norm, and $H(w)$ is the Shannon entropy: $H(w) = -\sum_{j=1}^k w_j \log w_j$.
Step 3) Given a cost matrix C as described above, the assigned class is

$$\arg \min_{\hat{y}} \sum_{j=1}^k w_j C(\hat{y}, Y_j(X)).$$

4. EXPERIMENTAL DETAILS

Each image was classified into one of the twelve classes by each classification algorithm. Due to the small ratio of labelled data (228 images) to number of classes (12), leave-one-out cross-validation was chosen to compare classification algorithms. For each sample X , classifier parameters (including the feature scalings calculated from CART) were estimated based on the other 227 sample points and the estimated class of X was determined using these parameters. The results shown in Table 2 and Table 3 are the average performance over all 228 images.

Images from the pipeline are overwhelmingly *normal*. The hand-labelled database is more evenly distributed over the classes. Using either density as a prior resulted in classifications skewed towards *normal*. The error rates reported in this paper are based instead on a uniform prior.

The consequences of different class mislabellings varied significantly. Failing to detect *cavities* may seriously compromise the future integrity of the pipeline, however the ramifications of confusing *osmosis* and *corrosion blisters* are much less significant. The misclassification costs were estimated by the researchers at NEO and this cost matrix was used in all classifiers except for the column titled LIME (0-1 cost) which was run with equal (0-1) costs for all classes to highlight the effect of NEO’s cost matrix. Misclassifying images of classes A (*normal*), E (*longitudinal welds*), H (*field joints*), I (*grinder marks*) and J (*MLF marks*) generally had small costs, whereas misclassifying images of classes F (*weld cavity*) and G (*welds too close*) generally had high costs. For the other classes (which we call medium cost classes) the cost varies significantly dependent on how the image is misclassified.

In Table 2 the average costs per class (calculated using NEO’s cost matrix) are listed for each algorithm. This information tells us that if there is an event of a particular class, the average cost is how much cost we expect to incur by the algorithm’s estimation. Given the small dataset and misclassification costs ranging from 5 to 600, a small number of serious errors causes a significant increase in average cost.

The different algorithms perform quite differently. MART and LIME were more strongly influenced by the cost matrix than the Gaussian methods (LDA and regularized QDA). In the low cost classes, listed above, LDA, regularized QDA

and LIME (0-1 cost) generally perform better than MART and LIME. Conversely, for most medium and high cost classes MART and LIME outperformed, often significantly, the other methods.

	MART	LDA	Reg. QDA	LIME	LIME (0-1 cost)
A	10.00	0.23	0.70	9.65	0.81
B	10.25	14.75	5.00	5.75	0.00
C	16.07	28.57	37.50	11.79	37.14
D	6.18	117.65	52.94	29.41	47.06
E	10.00	2.50	0.00	1.50	1.25
F	6.32	99.47	17.37	9.47	15.79
G	62.50	161.88	50.00	25.00	74.69
H	20.00	0.00	2.50	25.00	5.00
I	22.00	0.75	0.25	20.75	1.25
J	13.85	3.08	1.54	20.77	1.92
K	36.36	27.27	72.73	20.91	18.18
L	15.33	78.00	41.33	6.00	46.67
Mean Cost	19.07	44.51	23.49	15.50	20.81

Table 2. Mean expected cost for an event of a given class

In Table 3 the recall per class is listed for each algorithm. Recall for class Y is the number of images that belong to class Y that were correctly labelled. This table clearly shows the recall dropping for low cost events when NEO's cost matrix is used in the LIME estimation instead of LIME with 0-1 costs. LIME classification using 0-1 cost has higher recall than LIME on all but two events (and significantly higher average recall). The misclassifications by LIME with 0-1 cost were expensive according to NEO's cost matrix, however. LIME misclassifies all samples from class J, but given the low cost of these misclassifications the algorithm still maintains a low average cost.

5. FUTURE WORK

More data of critical events, such as *welds too close*, could alter the balance of the results and lead to more robust comparisons. The twelve class problem can be augmented with the simpler problem of separating 'normal' from 'abnormal' images. The abnormal images could be fed to a human discriminator.

Lower expected costs may be obtainable with a hybrid classifier incorporating the opinions of a number of different classification methods [6]. For example, a majority rule classifier might use MART, regularized QDA and LIME. If two of the classifiers agree to the class then the majority class is chosen. If no two classifiers agreed, then the LIME result could be used (with the assumption that although it might be wrong, it will provide on average the least costly misclassification).

	MART	LDA	Reg. QDA	LIME	LIME (0-1 cost)
A	0.49	0.98	0.91	0.44	0.91
B	0.45	0.90	0.95	0.65	1.00
C	0.07	0.64	0.50	0.21	0.50
D	0.76	0.53	0.82	0.88	0.82
E	0.70	0.90	1.00	0.85	0.95
F	0.79	0.74	0.89	0.68	0.95
G	0.69	0.63	0.69	0.75	0.75
H	0.60	1.00	0.95	0.50	0.90
I	0.25	0.90	0.95	0.15	0.85
J	0.23	0.69	0.85	0.00	0.77
K	0.36	0.82	0.36	0.73	0.82
L	0.20	0.00	0.33	0.47	0.33
Mean Recall	0.47	0.73	0.77	0.53	0.80

Table 3. Mean recall for an event of a given class

6. ACKNOWLEDGEMENTS

The authors would like to thank Norsk Elektro Optikk for providing data, as well as engineering and financial support. The authors also thank Richard Olshen for useful discussions.

7. REFERENCES

- [1] D. L. Atherton, "Magnetic inspection is key to ensuring safe pipelines," *Oil and Gas Journal*, vol. 87, no. 32, pp. 52–61, 1989.
- [2] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Trans Systems, Man and Cybernetics*, vol. 3, no. 6, pp. 610–621, 1973.
- [3] Leo Breiman, Jerome Friedman, Richard Olshen, and Charles Stone, *Classification and Regression Trees*, Chapman and Hall, United States of America, 1984.
- [4] Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning*, Springer-Verlag, New York, 2001.
- [5] Maya Gupta, "An information theory approach to supervised learning," *Stanford University Dissertation*, 2003.
- [6] F. Provost and T. Fawcett, "Robust classification for imprecise environments," *Machine Learning*, vol. 42, no. 3, pp. 203–210, 2001.