Pipeline Anomaly Assessment: Classifying Gastrointestinal Diseases by their Texture with a Novel Multimodal Classifier*

Craig L Champlin[2]

***Abstract*— We develop a multimodal classifier using a regularized sparse optimization approach with a modality grouping norm and test it against a dataset of gastrointestinal diseases which serve as a proxy for pipeline afflictions. With a 6 modality feature vector, we did 24% better than an SVM.**

I. Prologue

The motivation for the work presented here is, as specified in the directions, directed at pipelines. While this may not appear to be the case it is true. Please allow me to explain.

My literature search for the other half of this exam, "Robot Localization in Space Constrained Environments" was exhaustive in the domain of robotic pipeline inspection. While I may not have cited everything in this domain, I believe I did uncover and skim close to every bit of literature available. As I worked I came to understand what characterizes the domain. The idea that inspection vehicles are constrained to linear 4-DOF motion along the pipe is an artificial constraint. I discovered many examples which require unconstrained motion of an inspector. One example is the inspection of dam overflow pipes which have massive diameters, on the order of 15 to 20 feet [1]. Another example is the use of a free-swimming AUV submarine to inspect water mains [2]. And finally, I encountered recent developments in the use of tiny robots which move in the motion-unconstrained, flexible cavities of the human body [3].

The goal of all in-pipe navigation is to be able to monitor anomalies and report-out when something needs to be repaired. This is as true for water mains as it is for colons. The tasks required to meet this objective are to be able to navigate to a known location within a long featureless environment, recognize when a target location has been reached via robust place recognition, and finally be able to closely characterize target anomalies so that their changes and severity can be tracked through time. It is a daunting. What all of these examples have in common with the traditional idea of pipeline inspection is that all environments are incredibly feature-poor. They are all long and narrow. They are all featureless. Every bit of information along the route, including anomalies, become important navigational aids. The domain includes oil pipes, tunnels, sewers, and body cavities – like colons.

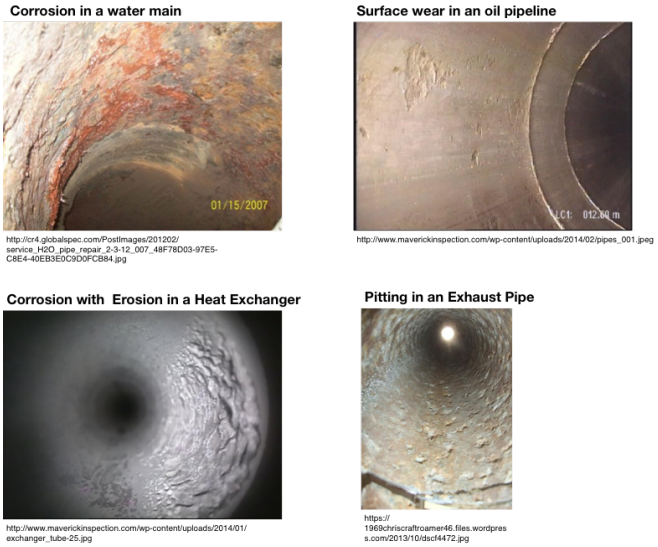

Fig. 1. Examples of different types of pipes and pipe "diseases". Compare these with the images of GI diseases on the next page. Both domains are characterized by ailments which appear as subtle textural differences.

So, when I couldn't find any datasets or significant papers discussing anomalies in pipes I found the emerging field of robotic inspection of colons [4] which are very similar to pipes. Compare Figure 1 with Figure 2 to see the similarities of the two domains. The requirements for anomaly detection and characterization actives in a colon are very similar to those in an oil or water pipe. While the details of the measuring modalities may change, the domains share bad lighting, obtuse camera angles, and difficulty segmenting healthy regions from trouble spots. Given the lack of availability of supporting research and datasets on infrastructure pipelines I have been forced to adapt the instructions of this exam to logically support the end-game of unconstrained pipe/tube/tunnel inspection by working with in-body pipelines.

I have written this report as if colon inspection was the original objective. I have tried to be true to the original literature requirements to make sure I have a wide mix of supporting papers. Rather than stick to arbitrary paper categories I selected categories of papers which support the work. I think you will see that I have more than met the spirit of the rule, which was to demonstrate the ability to discriminate good papers from bad and to apply current published research to a novel goal. You will find the contents of my literature review in the *Related Work* section of this report. To help you identify if I have met the quantity

*This work was prepared as a requirement of the Computer Science PhD qualifying exam.

[1]Craig Champlin is with the Human-Centered Robotics Laboratory in the College of Engineering and Computational Science (CECS), Colorado School of Mines, 1400 Illinois St, Golden, CO 80401, USA cchampli@mines.edu

of papers required, I have placed a table in the *Appendix* which illustrates the adapted minimum requirements of the examination instructions. I only include a reference in this list if I met the required standard of including a *comparison of the pros and cons* with respect to the project.

Please forgive my bending of the rules, but it is my understanding that the intent of this exam is to demonstrate that I am capable of performing independent research. After examining my work, I hope you will agree that I am capable.

Thank you.

II. Introduction

Colorectal cancer is the third most common cancer diagnosed in both men and women, and the second leading cause of cancer death in the US, which is expected to cause about 49,700 deaths during 2015 [5]. Early detection and diagnosis of colonic diseases could prevent the majority of the deaths, since most such cancers develop from adenomatous polyps, which can be identified and removed. When a colorectal cancer is detected at its earlier stages, the patient?s survival rate can also be significantly increased, to approximately 90

In recent years, non-invasive colon screening technologies known as capsule endoscopes (CE)s have attracted wide attention, due to their potential to reduce the risks and expense of conventional endoscopy. Robotic Capsule Endoscope (RCE)s have also been introduced with the hope that these actively driven devices may one day completely replace conventional endoscopes [3]. For example, the PillCam can help doctors screen the colon, using a micro video camera contained in a disposable capsule that is able to pass through the human GI tract naturally [6]. Robotic capsules were also introduced [7]. A novel RCE was developed in our previous research [8] to enable painless and patient-friendly GI tract investigation, which is able to navigate in the human GI tract and actively observe tissues of interest [8]. It is envisioned in the future, these non-invasive technologies would allow patients to perform regular GI tract screening and diagnosis at home.

These systems generate lots of data. Medical professionals can only process a few images at a time. There is a need to provide these professionals assistance by developing computer algorithms which can provide preliminary diagnostics to thousands of images while returning a few interesting results for the professional to review [9]. Additionally the performance of human evaluators varies by the individual and by time of day. Developing a computer aided diagnostics system would help provide stability and continuity to the diagnostic process. According to a 2017 bibliographic survey of computer aided diagnostics there has been no work done in the domain of computer classification of GI diseases. According to this work the domain is *nascent stages of development* [4]. Any results we provide will be welcome.

Here we will concern ourselves with the multimodal classification of gastrointestinal (GI) diseases. We are going to explore the effectiveness of using texture to classify GI diseases from segmented images. As a test-case we will use a series of endoscope images which were captured using

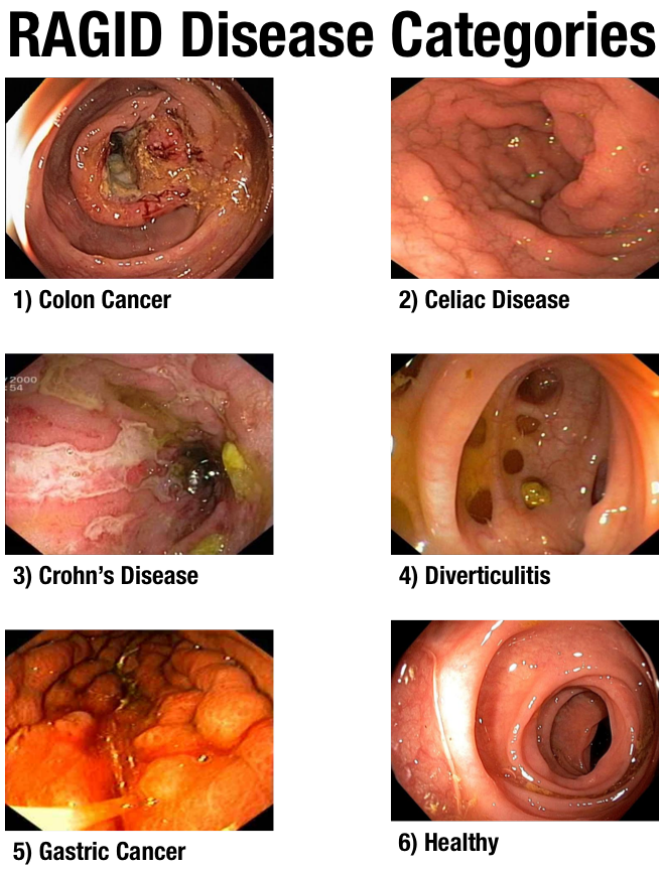

Fig. 2. The six categories from the RAGID dataset we used for testing our algorithm. This set consists of approximately 30 RCE images per category, for a total of 181 images. The disease portion of each image was segmented with a convolutional neural network in [10]. We will train on the segmented portions of these images.

an RCE. These images in the RAGID dataset have been labeled and have been segmented so that we have a mask which locates the diseased region [10]. See Figure 2 for samples of this dataset. To classify these images we develop a novel multimodal classifier based on a sparse optimization formulation which relies on two regularization terms, one of which associates features with their modality of origin. We call our classifier MC2T for *Multimodal Classifier with 2 Terms*. We evaluate the performance of our classifier against a support vector machine using a rbf kernel.

We make three contributions in this paper:

1) We introduce a region-based approach for classifying anomalous areas of medical imagery based on texture.
2) We develop a novel multi-modal classifier which learns the discriminative capabilities of each modality.
3) We performed a comparative analysis of feature descriptors for GI disease classification based on texture.

The rest of this paper is organized as follows. In section II we describe related research on two topics (1) classification of medical anomalies and (2) multimodal classifiers. In section III we introduce the MC2T approach to multimodal learning. In section IV we present our classification of gastrointestinal diseases using MC2T. Finally we have a concluding discus-

sion about the implications of our work and future directions it should take.

III. Related Work

Our goal here is to develop a multimodal classifier which can diagnose medical anomalies in the GI system. As such we will review key papers in the domains of medical anomaly classification and the more general domain of classification of image regions using multiple modalities. Since we are using regularized optimization for this task, we will pay particular attention to works which use these techniques.

As to the requirements of the qualifying exam, this section constitutes my literature review. For the benefit of a scoring rubric there is a tally of papers and their classification in the *Appendix*.

A. Medical Anomaly Classification

The field of medical anomaly classification is vast and extremely active. According to Takahashi in their 2017 bibliographic analysis of the field, the general term is *computer aided diagnosis* or CAD and it was originally developed for breast cancer using mammography in the 1960's [4]. In this field the computer assists medical professionals by performing disease detection and diagnosis from medical imagery. While there is significant work in the field, out of 19 references surveyed all were written using CT scans for CAD and none of them addressed direct imaging within the colon. While this means the field is wide open for research, it also means that there is little direct research to lean on. There are two primary questions we need answered from the literature, we need to know how medical anomalies are classified and what feature descriptors are available for assessment of optical imagery of tissue.

A 2014 study by Mamonov specifically examined the problem of classifying imagery from RCEs [11]. The authors observed that RCE imagery is circular and that this edge can cause artifacts when classifying RCE images. They responded by extrapolating image values into the non-exposed region of the image. We will respond by operating only on a segmented region and by making sure that the segmented region does not extend into the unexposed frame boundary. Mamonov has identified texture a being "an important first step of the algorithm". We will follow this guidance an ensure that our feature descriptors are free from geometric information. We will elect to focus on textural information. As a textural descriptor the authors used multiple iterations of convolution with a gaussian kernel to separate the image into textural and geometric components. They threshold the textural component and discard a frame prior to classification if the textural values fall outside of these preset boundaries. This is to eliminate healthy tissue and images which may be corrupted by bubbles or other debris. Outside of this textural analysis, Mamonov creates a set of geometric descriptor for classification. Upon observing the images in our dataset, geometric descriptors may work for specific diseases. As a first pass in a general classification scheme, however, this may be too specific.

K Suzuki put together two comprehensive review of machine learning for medical diagnosis. These studies were 15 years apart. The first study in 2012 explored features for diagnosis of colon polyps with traditional endoscopes [12]. The second study in 2017 is a survey of computer aided diagnostics for lung cancers [13]. The surveys are worth mentioning for two reasons. First they are both comprehensive, the 2012 study cites 25 studies which have done machine based classification. The 2017 study mentions features 56 times. For each study in both surveys the author mentions the feature descriptor used, the classification technique, and the accuracy the method achieved. Secondly, in both cases there is a notable absence of typical features we see in the computer-science machine learning domain. Typical features such as SIFT, and HOG are notable by their absence. Most features are geometric - such as topographic height, diameter, curvature indexes and shape indexes. It would seem that this approach would be prone to overfitting as they seem very specific. Perhaps there is not a lot of collaboration with computer scientists in these fields?

Recent work by Biswas in 2016 appears to have achieved remarkable success at classifying distinguishing from normal 4 diseases of the colon: polyps, ulcers, Chron's disease, and sigmoid colon [14]. With a group of 1,185 labeled images, they were able to achieve 98% accuracy with a 20-fold cross validation holding 150 test images back per fold. Their multi-step approach replaced traditional segmentation/feature extraction schemes. Instead this group processed whole colon images through three steps prior to classification. First they performed a radon transform, which creates a radial set of intensity slices through the image from the center. This is their starting representation of the image. They assemble these slices into a single vector. Next they apply a cross wavelet transform using a Morlet wavelet as the mother wavelet to this vector. This results in a square matrix. As a final step before classification they extract the principal components from this matrix and take (apparently) the first principal component as the feature vector for training with a support vector machine. This seems extraordinary. I would like to try this technique as a benchmark.

B. Multimodal Classifiers

In the past several years multimodal classification has been intensively investigated in multiple domains because of its importance in clearing up the problem of perceptual aliasing. Some of the variety of domains taking this approach this are robot perception [15], robust loop closure [16], remote sensing [17], and affective computing [18], among others. One of the key questions to answer when combining multiple sensing modalities is how to combine modalities such that the most discriminative classifiers have the most influence on the final result. One approach is to combine modalities

Classification techniques which rely heavily on a single modality are subject to perceptual aliasing. Adding another sensing modality to the system can often resolve the problem. The multi-modal classification of anomalous regions is a general task with wide applicability to domains as diverse as

medical diagnosis, infrastructure maintenance, and land use analysis. The automatic classification of anomalies replaces subjective human assessment with repeatable machine-based evaluations. The goal of this paper is to support a reduction in perceptually aliasing by developing a new multimodal anomaly classification technique.

The multimodal classification of the diseased regions of the GI track is an example of the general image region classification problem. The steps involved in classifying an image region are as follows:

1) preprocess the image
2) segment regions
3) generate features
4) classify regions

This process, as described, does not consider multimodality. In a multimodal classification scenario when and where we fuse modes together will impact the way the classification system behaves. The survey paper by Gomez-Chova et al on multimodal classification of remote sensing images presents a taxonomy of where in the process fusion of modalities can occur [17]. According to this taxonomy there are thee places where we can fuse information from different modalities: (1) at the *pixel level* by fusing modalities into a single image, (2) at the *feature level* by creating unique features for each mode, (3) and at the *decision level* by creating different classifiers for each modality.

Our work is best characterized as being applied at the *feature level*. We create unique features for each modality and combine these features into a single classifier which learns the discriminative capabilities of each mode. We use a modal grouping term which insures that our system learns discriminative modalities, not just discriminative features. This term also insures that modes with high dimensionality do not overwhelm smaller ones.

A 2016 paper by Kong evaluated a wide range of breast cancer features for their ability to predict the recurrence or no-recurrence of the disease [19]. To perform this assessment the group observed that linear discriminant analysis, which aims to maximize the ratio between interclass variability (between) and intraclass variability (within), becomes indeterminate when the number of feature dimensions is larger than the number of samples, a common problem in domains like medical diagnostics where data is hard to obtain. In response, Kong adapted a related approach called Maximal Margin Criterion which formulates the variability ratio objective function as an eigenvalue problem which searches to find the U which maximizes the difference between the two variabilities. To this basic approach, Kong added a $\ell_{2,1}$ norm regularization term which they demonstrate insures the optimization problem converges to a local optimum. This approach is similar to ours except our objective function looks at projection error instead of intraclass variability.

This next paper, a 2017 paper by Latif et al, has been accepted by Robotics and Autonomous Systems [16] but has not yet been published is interesting on multiple levels. The premise of the paper is that it uses sparse optimization for robust loop closing. We are using a sparse optimization technique here, which is why it is included. Since I am generally interested in loop closing for pipeline inspection, it caught my eye from that perspective also. The authors present a unique take on an ℓ_1-minimization problem. They formulate loop closure as a redundant minimization problem for which "a sparse solution is sought for a given observation." In a sense, they use the approach to look for the best match to our feature vector, X in a dictionary matrix, $B = [b_1, ...b_m]$ in the case of loop closure, the dictionary is a set of reasonable location guesses which are in proximity of the most recent known location. To apply this to colon disease diagnosis, this dictionary could be a most-characteristic heterogeneous descriptor for a given disease. Interestingly, this is exactly what we do with our W matrix. We learn our most-characteristic heterogeneous descriptor during training. Our W matrix is their B matrix. During classification we compare a test-case against this vector to find the one which fits the best. See the development of our technique later in this document.

C. Shared Representative Appearance Learning

Our work here leverages an in-press paper by a member of our group, Fei Han [15]. In this work, called SRAL for *Shared Representative Appearance Learning*, Han has the goal of evaluating the power of individual observation modalities to discriminate a scene under a set of widely variable seasonal conditions. The goal is to learn weightings for each modality which minimize the feature-space difference score between two identical locations.

To accomplish this goal the authors use a sparse optimization formulation whereby the objective function is the Frobenius norm of the difference between the projected feature representation and the equivalent low dimensional label space. The authors use a novel, modally-biased relaxation term to group features from each mode together during training. This so-called ℓ_M norm term provides intragroup structure across the features within a modality. Additionally the group's technique applies a more common $\ell_{2,1}$ norm to reduce noise and ensure convergence. Since we use the SRAL formulation as the starting point for our own work, it is worth repeating here.

The basic structure of the approach is a regularized sparse optimization problem. If we take m to be the number of modalities, then $d = \sum_{j=1}^{m} d_j$ is the total number of row elements in the heterogeneous multimodal feature vector, $X = [x_1, ..., x_n] \in \mathbb{R}^{dxn}$, where d_j is the length of the feature vector for modality j, n is the number of observations, and $x_i = \left[\left(x_i^1\right)^\intercal, ..., (x_i^m)^\intercal\right]^\intercal \in \mathbb{R}^d$ is a heterogenous feature vector for a single observation, i. Now, if we define c to be the number of scene conditions we need to recognize then $Y = [y_1, ..., y_n]^\intercal \in \mathbb{R}^{nxc}$ represents the ground truth label of an image belonging to a particular location under a particular set of scene condtions.

Our goal is to learn the weighting matrix, $W \in \mathbb{R}^{dxc}$ which maps the higher dimensional feature vector, X to the lower dimensional label space, Y. We find W as the solution to minimization problem (1).

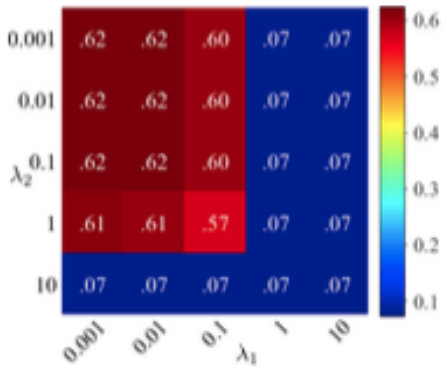

Fig. 3. A sensitivity analysis to study the effect of the weighting terms λ_1 and λ_2 on the place recognition performance in [20].

$$\min_W \left\|X^T W - Y\right\|_F^2 + \lambda_1 \|W\|_{2,1} + \lambda_2 \|W\|_M \tag{1}$$

The terms of Equation 1 can be described as follows:

- $\left\|X^T W - Y\right\|_F^2$
  - error term between the prediction and ground truth
  - when this is zero our prediction is an exact match
- $\lambda_1 \|W\|_{2,1}$
  - $\ell_{2,1}$-norm which enforces sparsity on rows of W
  - this denoises the result and drives us to a single answer
  - this term also drives the answer to having similar weights for each feature *(row in X)*
- $\lambda_2 \|W\|_M$
  - ℓ_M-norm synchronizes discriminative powers by mode
  - enforces feature membership in a group *(block of rows in X)*
  - helps balance the importance of small and large modes
  - also imposes sparsity between modes

We calculate the two norms as follows:

$$\|W\|_{2,1} = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{c} w_{ij}^2} = \sum_{i=1}^{d} \left\|w^i\right\|_2 \tag{2}$$

and,

$$\|W\|_M = \sum_{i=1}^{m} \sqrt{\sum_{p=1}^{d_i} \sum_{q=1}^{c} \left(w_{pq}^i\right)^2} = \sum_{i=1}^{m} \left\|W^i\right\|_F \tag{3}$$

The terms λ_1 and λ_2 are weighting terms that allow us to tweak the influence of the two regularization terms. In [20] Han does a sensitivity analysis search for the value of these weight terms, shown as Figure 3 and used the values $\lambda_1 = \lambda_2 := 0.1$. An open question is if we need to adjust these parameters here for MC2T.

The regularization terms in Equations (2) and (3) are not smooth and so they prevent us from finding the minimum of Equation (1) analytically. Han devised the gradient descent algorithm shown in Figure 4 to solve it. The objective

Input : $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_n] \in \Re^{d \times n}$ and $\mathbf{Y} = [\mathbf{y}_1, \cdots, \mathbf{y}_n]^\top \in \Re^{n \times c}$

1: Let $t = 1$. Initialize $\mathbf{W}(t)$ by solving $\min_{\mathbf{W}} \|\mathbf{X}^\top \mathbf{W} - \mathbf{Y}\|_F^2$.
2: **while** *not converge* **do**
3: Calculate the block diagonal matrix $\mathbf{D}(t+1)$, where the i-th diagonal element of $\mathbf{D}(t+1)$ is $\frac{1}{2\|\mathbf{w}^i(t)\|_2}\mathbf{I}_i$.
4: Calculate the block diagonal matrix $\tilde{\mathbf{D}}(t+1)$, where the i-th diagonal block of $\tilde{\mathbf{D}}(t+1)$ is $\frac{1}{2\|\mathbf{W}^i(t)\|_F}\mathbf{I}_i$.
5: For each $\mathbf{w}_i (1 \le i \le c)$, $\mathbf{w}_i(t+1) = \left(\mathbf{X}\mathbf{X}^\top + \gamma_1 \mathbf{D}(t+1) + \gamma_2 \tilde{\mathbf{D}}(t+1)\right)^{-1} \mathbf{X}\mathbf{y}_i$.
6: $t = t + 1$.
7: **end**

Output: $\mathbf{W} = \mathbf{W}(t) \in \Re^{d \times c}$

Fig. 4. The gradient descent algorithm developed in [20]. The objective function in step 5 is the result of setting the derivative of the error function in Equation (1) with respect to w_i to zero and then solve for w_i which is the term we wish to find.

function in step 5 is the result of setting the derivative of eqn (1) with respect to w_i to a zero vector and then solving this for w_i which is the term we wish to find. The D terms are summary terms for the complex diagonal terms which arise when we take the derivative of the $\ell_{2,1}$ and ℓ_M norms. They are defined in steps 3 and 4 of the algorithm. To solve, we iterate until w_i converges to a threshold value.

The authors use their newly found weight matrix, W, to find a concatenation weight for each modality. These weights reflect the ability for each mode to discriminate a scene across widely variable conditions. To use this for place recognition they weight each mode according to the Frobenius norm for each modality in the weight matrix. This gives them a single value per modality. In use, they normalize these weights on the interval $[0,1]$ and and multiply each modality in each feature by its corresponding weight. Then they are able to compare a test image against a database image by calculating a weighted sum of absolute difference score. This score can then be used in the place recognition algorithm of your choice.

One nice feature of this approach is that it is possible to know which modes are not able to discriminate within a particular environment. These are the modalities which have low weights. In practice then a user could not collect this data, saving time and energy on something which would be demonstrably unnecessary.

IV. MC2T Algorithm

Recall that our objective is to develop a *feature-level* multimodal classifier as defined by Gomez-Chova [17]. In the domain of pipes, where cameras are not the best way of looking around, we wish to combine whatever modalities we have available. It is rare to have video inside a pipe, but it is not uncommon to have magnetic flux loss readings, ultrasound readings, and eddy current readings. It is our hope that someday we can use these alternate modalities to distinguish construction features from microbial-induced corrosion form

weld pitting from mechanical scouring and other pathologies. This is a multimodal classification problem. It is why we are here.

Our *Multimodal Classifier with 2 Terms*, MC2T algorithm is very similar to the SRAL algorithm developed by Han [20] as presented in the previous section. While Han used the weight matrix W for a direct comparison of feature scores we will formulate the problem a little differently.

Let:

- c be the number of categories we wish to classify
- $d = \sum_{j=1}^{m} d_j$ be the sum of the lengths of all modalities, s.t.
  - d_j is the length of the feature vector for measurement modality j
  - and m is the number of measurement modalities
- n be the number of observations
- so that, the heterogeneous feature vector is
  - $X = [x_1, ..., x_n] \in \mathbb{R}^{dxn}$, with feature vectors
  - $x_i = \left[\left(x_i^1\right)^\intercal, ..., \left(x_i^m\right)^\intercal \right]^\intercal \in \mathbb{R}^d$ for observation i

Then, then $Y = [y_1, ..., y_n]^\intercal \in \mathbb{R}^{nxc}$ represents the ground truth assignment of a training image to its proper category. Our goal is to learn the $W \in \mathbb{R}^{dxc}$, the weight matrix. To accomplish this, we use the algorithmic approach from [20] with this different formulation.

Once we have W we can use it to classify an unknown feature vector by $Y_{guess} = x_{test}^\intercal W$ where x_{test} is the unknown feature vector.

V. Evaluating the MC2T Algorithm

Absent good mechanical survey data on classical pipelines, we will use the multimodal classification of GI diseases as a proxy for our target objective. It is a good proxy since, as evident in Figure 2 which shows the GI disease categories we are considering, distinguishing healthy from diseased conditions is a subtle distinction in texture. This is similar to the pipe "diseases" seen in Figure **??**. Bad conditions are not a simple as a dark splotch on a light background. This would be easy to segment and detect, instead categories $\{2,4,6\}$ down the right side of the figure are very similar in color and texture. The conditions in a pipeline are similarly subtle.

Here we will consider how best to classify these diseases without regard for the classifier we use. Having established a good machine learning approach, we will use this approach with our MC2T multimodal classifier with a Support Vector Machine as a control. This comparison will take place in the next section.

A. Modality Selection

Since our dataset is strictly visual, a single mode, and the classifier we wish to test is multimodal we will use visual descriptors as proxies for multimodal measurement. This is a good substitution since when measuring a scene with multiple modes all modes will represent the underlying structure. Each representation will differ in size and resolution and significant features. Calculating visual descriptors will simulate this relationship between a modality and the underlying structure. See Figure 5 for an example of some of the descriptors we will use. All four descriptors in this image were drawn from the same region in the underlying source image. The first column represents the base visual representation from which the descriptor in the center column is based. The right column shows the descriptor name and the size of its feature vector.

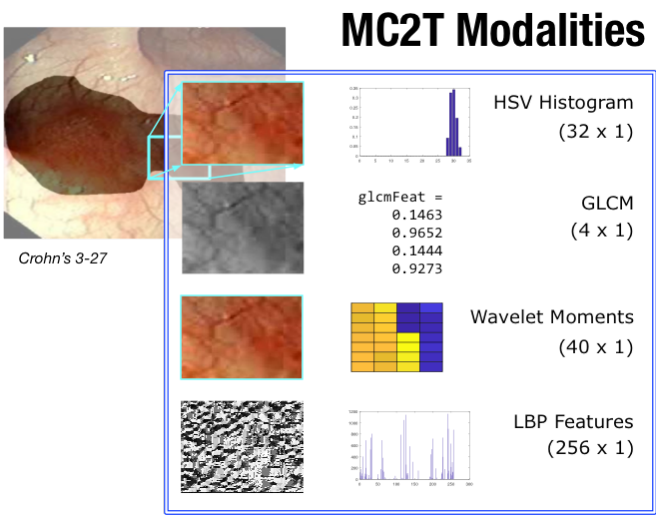

Fig. 5. Example of the descriptors used as sensing modalities for this work. These descriptors were all drawn from the same patch on image 3-27, which represents Crohn's disease. The first column is the source patch for the descriptor. The second column is the descriptor resulting from this patch. The final column is the descriptor name and the size of the resulting feature vector.

Our literature search revealed that much medical disease classification in CAD is done using textural and hand-crafted geometric descriptors. Since we do not have to expertise to craft medically significant geometric descriptors, we will attempt to characterize each GI disease using features which represent textures. Unlike medical practitioners who will are interested in precision diagnostics, we are interested in evaluating our classifier's ability to leverage the information from multiple modalities. To that end, we will attempt to classify using ONLY textural information. Since we want to avoid encoding geometric relationships into our learning we will avoid traditional geometric descriptors which are good at representing shapes and points in images. Examples of descriptors we do NOT want to use include SIFT, GIST, and Harris corners. Whole image descriptors, as in the ones used for SeqSLAM are also too much tied to the geometric arrangements of items in a scene. We will avoid those also. To further exclude scale and geometric information from our learning, we break our images into patches as we will describe in the *data pipeline* section below.

For our evaluation of MC2T we will use the descriptors as shown in Table I.

B. Data Pipeline

As mentioned in the previous section, since we are interested in evaluating our classifier and not so much interested in diagnosing diseases, we are interested in the challenge

TABLE I
DESCRIPTORS AS MODALITIES

Name	Len	Description
GLCM	4	gray level co-occurance parameters [21]
color	6	statistical representation of color content [22]
hsv	32	histogram of intensity values per image plane[1]
wavelets	40	first two wavelet moments[2]
HOG	186	Histogram of oriented gradients [23]
LBP	256	Local Binary Pattern[3]

of a purely textural representation of the RAGID dataset. As such, we wish to remove as much geometric and scale information from our training as possible. To that end, we have elected to split the segmented region of each image into patches and train on those individual patches. This also gives us significantly more training data, since we are training on image patches instead of whole or segmented images.

Figure 6 illustrates the data pipeline we used. The raw image in part 1 as it exists in the RAGID dataset. In step 2 we apply the segmentation mask which Amabarak [10] learned with a CNN. We overlay this with a 10x10 grid as shown in part 3. If a grid cell has at least 40

This is where the training and testing pathways diverge. If we are training, we add all the patches from all images into the heterogeneous feature matrix, X, and each patch gets its own entry in the ground truth matrix, Y. Once we have processed all patches on all images, we run the algorithm in Figure **??** until it converges to find the weight matrix, W.

On the other hand, if we are testing an unknown image we multiply each patch from the image against our learned W to find a guess for that patch. This is what is shown in step 5 of Figure 6. Each patch cell gets its own set of guesses as shown by the bar graph of category scores in each cell. We take the maximum of each patch category score as the category for that patch. This is shown in step 6 as the numbers in each cell. Each number corresponds to a disease category as identified in Figure **??**. The final step for classifying an unknown image is to aggregate the scores from each patch. This is shown as the histogram in the bottom right corner of the figure. This histogram represents the count of scores for each category. In this particular example, while there were many votes for class 2, celiac disease, most of the votes were for class 5, gastric cancer. We record the final score of this image as 5, gastric cancer.

Note: A quick glance at these two classifications in Figure **??** shows how the texture of these two categories is strikingly similar. This is a reasonable point of confusion for the algorithm.

VI. EXPERIMENTAL RESULTS

We trained our MC2T classifier against 181 images in the RAGID dataset. We break each image into a 10x10 grid of cells and generate a patch for training or testing whenever a cell is mostly occupied with the segmentation mask. On average about 30% of those patches are occupied. This gives us around 5,500 patch images for training. Additionally, we

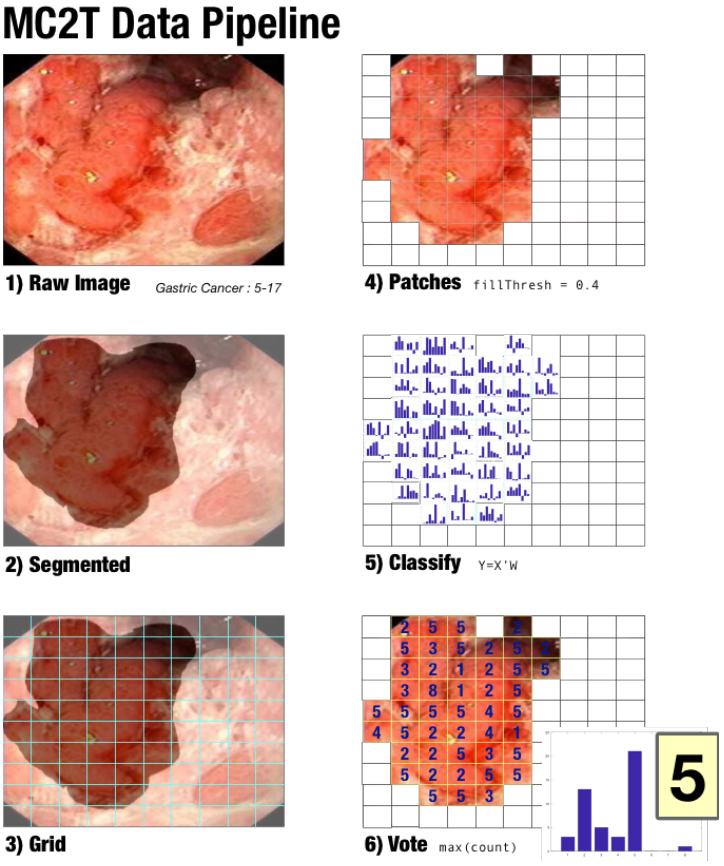

Fig. 6. The MC2T data pipeline. The weight matrix, W, is trained against patches. For classification the unknown heterogeneous feature vector, $x_{unknown}$ is multiplied by the learned W to extract a guess, Y_{guess}. The class for each patch is taken as the highest guess. The identity of the reassembled image is the most frequent guess.

perform 10-fold cross validation. We feel that there is enough training data for this test of our classifier to be effective.

To eliminate feature selection from the equation, we experimented with the number of cells and the threshold of how much of a cell to be occupied. We achieved a classification result maximum with a 10x10 grid and a fill threshold of 40

Next we experimented with individual modalities (descriptors). For each experiment we set grid size and fill threshold to the values we established earlier. We ran a 5-fold cross validation with a single modality. The classification results are shown in the list below as the overall positive accuracy. This is the number of true positives over the total number of trials during the 5-fold cross validation. It is a pessimistic measurement per [24] who prefers the macro-aggregated F1 score for this type of analysis. However this is a harder to understand metric which incorporates both precision and the recall metrics. It is much quicker to grasp the accuracy as a quick evaluation tool.

Classification accuracy – Single modality – Increasing Size

- 0.187 – 004 elem – **glcm**
- 0.195 – 006 elem – **color**
- 0.166 – 032 elem – **hsv**

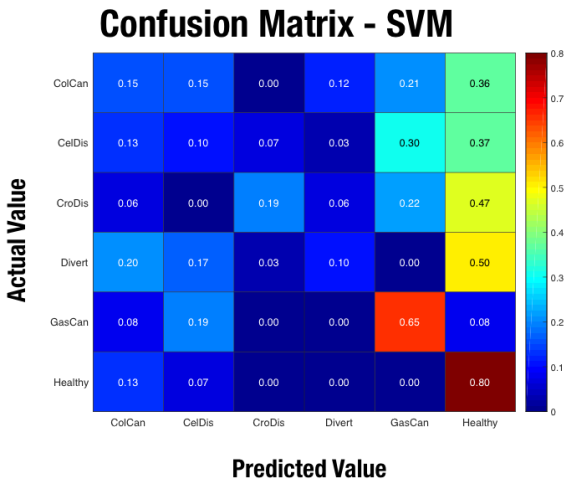

Fig. 7. The confusion matrix from running all modalities through 10-fold cross validation with a SVM classifier.

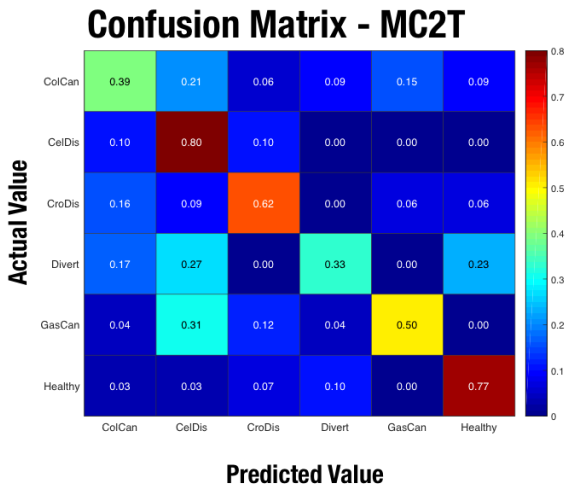

Fig. 8. The confusion matrix from running all modalities through 10-fold cross validation with our new MC2T classifier.

- 0.212 – 040 elem – **wavelet**
- 0.241 – 186 elem – **HOG**
- 0.349 – 256 elem – **LBP**

It is interesting to note that there is a correlation between the size of a modality and its ability to discriminate for classification. Based on this result, we should see an improvement in discrimination capability from simply concatenating these modalities into composite feature vector for classification. This is simply because the increased discriminative capability of the higher dimensionality vector. We will control for this phenomenon by comparing our multimodal classifier against a support vector machine. The support vector machine results will reflect the additional descriptive capability introduced by the longer feature vector. We can then assess the ability of our classifier to incorporate modal information.

We ran this experiment next. For this set of experiments, we did a 10-fold cross validation. We used a feature vector which was a concatenation of all 6 modalities listed above. All parameters were identical other than the classifier. These are the classification results:

- 0.320 - SVM
- 0.569 – MC2T

It was somewhat unexpected that the SVM did not show improvement above the classification capability of solitary LBP descriptor. This is most likely because we did not tune the SVM by finding its hyperparameters. We know that this will improve its performance. It would be interesting to see if this is able to close the 24% gap in accuracy between the two classifiers.

Finally we present the confusion matrices for each of these two experiments. These are shown as Figure 7 and Figure 8. While the classification accuracy of neither method is super fantastic, this is likely has to do with the modalities we selected. What is interesting is that there is a clear pattern here. The SVM misclassified many items as being healthy. while the MC2T did much better at correctly identifying diseases.

While they are not shown here, we did calculate the precision and recall metrics for each matrix as well as other parameters as suggested by [24] .

VII. Discussion

Is the better performance of the multimodal MC2T classifier because our technique does a good job of taking advantage of the information content of different modalities? Or, is it because we didn't optimize our support vector machine. The preliminary results are interesting and warrant further investigation.

Clearly optimizing the SVM is something I would like to try. Other things things to try in the future include experimenting with the regularization terms. I would like to try with and without each term as well as experiment with different weightings. I would also like to compare this multimodal technique with the convolution SVM kernel mentioned by Mamonov in [11].

VIII. Appendix

This appendix classifies the papers found during the literature search on pipeline anomaly detection. This search reflects oral instructions to adjust the focus of the literature search to cover medical anomaly classification vis a vis pipeline anomaly detection.

TABLE II
Literature Search Summary

Requirement	Citations
Reputable survey paper	Takahashi:2017 [4]
Medical anomaly classification (3)	Mamonov:2014 [11] Suzuki:2017 [13] Biswas:2016 [14]
Multimodal anomaly classification (4)	Gomez-Chova:2015 [17] Han:2017 [15] Kong:2016 [19] Latif:2017 [16]

ACKNOWLEDGMENT

Forrest and Ansel, this is for you.

REFERENCES

[1] T. Özaslan, K. Mohta, J. Keller, Y. Mulgaonkar, C. J. Taylor, V. Kumar, J. M. Wozencraft, and T. Hood, "Towards fully autonomous visual inspection of dark featureless dam penstocks using mavs," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct 2016, pp. 4998–5005.

[2] D. Krys and H. Najjaran, "Development of visual simultaneous localization and mapping (vslam) for a pipe inspection robot," in *2007 International Symposium on Computational Intelligence in Robotics and Automation*, June 2007, pp. 344–349.

[3] J. C. Norton, "The design and development of a mobile colonoscopy robot," Ph.D. dissertation, University of Leeds, 2017.

[4] R. Takahashi and Y. Kajikawa, "Computer-aided diagnosis: A survey with bibliometric analysis," *International Journal of Medical Informatics*, vol. 101, pp. 58 – 67, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1386505617300357

[5] American Cancer Society, "Cancer facts and statistics," [Online; accessed 23-April-2017]. [Online]. Available: http://www.cancer.org/research/cancerfactsandstatistics/

[6] A. Sieg, K. Friedrich, and U. Sieg, "Is pillcam colon capsule endoscopy ready for colorectal cancer screening[quest] a prospective feasibility study in a community gastroenterology practice," *Am J Gastroenterol*, vol. 104, no. 4, pp. 848–854, 02 2009. [Online]. Available: http://dx.doi.org/10.1038/ajg.2008.163

[7] P. Valdastri, R. J. W. III, C. Quaglia, M. Quirini, A. Menciassi, and P. Dario, "A new mechanism for mesoscale legged locomotion in compliant tubular environments," *IEEE Transactions on Robotics*, vol. 25, no. 5, pp. 1047–1057, Oct 2009.

[8] L. J. Sliker, M. D. Kern, J. A. Schoen, and M. E. Rentschler, "Surgical evaluation of a novel tethered robotic capsule endoscope using micro-patterned treads," *Surgical Endoscopy*, vol. 26, no. 10, pp. 2862–2869, 2012. [Online]. Available: http://dx.doi.org/10.1007/s00464-012-2271-y

[9] S. Zhang and D. Metaxas, "Large-scale medical image analytics: Recent methodologies, applications and future directions," *Medical Image Analysis*, vol. 33, pp. 98 – 101, 2016, 20th anniversary of the Medical Image Analysis journal (MedIA). [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1361841516300883

[10] A. Ambarak, J. Prendergast, M. Rentschler, and H. Zhang, "Toward automated detection of gastrointestinal diseases using deep neural networks," 2015, colon anomaly classification.

[11] A. V. Mamonov, I. N. Figueiredo, P. N. Figueiredo, and Y. H. R. Tsai, "Automated polyp detection in colon capsule endoscopy," *IEEE Transactions on Medical Imaging*, vol. 33, no. 7, pp. 1488–1502, July 2014.

[12] K. Suzuki, "A review of computer-aided diagnosis in thoracic and colonic imaging," *Quantitative Imaging in Medicine and Surgery*, vol. 2, no. 3, 2012. [Online]. Available: http://qims.amegroups.com/article/view/1077

[13] ——, *Computer-Aided Detection of Lung Cancer*. Singapore: Springer Singapore, 2017, pp. 9–40. [Online]. Available: http://dx.doi.org/10.1007/978-981-10-2945-5_2

[14] M. Biswas, A. Bhattacharya, and D. Dey, "Classification of various colon diseases in colonoscopy video using cross-wavelet features," in *2016 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, March 2016, pp. 2141–2145.

[15] F. Han, X. Yang, Y. Zhang, and H. Zhang, "Sequence-based multimodal apprenticeship learning for robot perception and decision making," *CoRR*, vol. abs/1702.07475, 2017. [Online]. Available: http://arxiv.org/abs/1702.07475

[16] Y. Latif, G. Huang, J. Leonard, and J. Neira, "Sparse optimization for robust and efficient loop closing," *Robotics and Autonomous Systems*, pp. –, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0921889015302281

[17] L. Gómez-Chova, D. Tuia, G. Moser, and G. Camps-Valls, "Multimodal classification of remote sensing images: A review and future directions," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1560–1584, Sept 2015.

[18] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98 – 125, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253517300738

[19] H. Kong, Z. Lai, X. Wang, and F. Liu, "Breast cancer discriminant feature analysis for diagnosis via jointly sparse learning," *Neurocomputing*, vol. 177, pp. 198 – 205, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S092523121501752X

[20] F. Han, X. Yang, Y. Deng, M. Rentschler, D. Yang, and H. Zhang, "Sral: Shared representative appearance learning for long-term visual place recognition," *IEEE Robotics and Automation Letters*, 2017.

[21] R. F. Walker, P. T. Jackway, and I. D. Longstaff, "Recent developments in the use of the co-occurrence matrix for texture recognition," in *Proceedings of 13th International Conference on Digital Signal Processing*, vol. 1, Jul 1997, pp. 63–65 vol.1.

[22] W.-Y. Ma and H. J. Zhang, "Benchmarking of image features for content-based retrieval," in *Conference Record of Thirty-Second Asilomar Conference on Signals, Systems and Computers (Cat. No.98CH36284)*, vol. 1, Nov 1998, pp. 253–257 vol.1.

[23] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *Proceedings of the 18th ACM International Conference on Multimedia*, ser. MM '10. New York, NY, USA: ACM, 2010, pp. 1469–1472. [Online]. Available: http://doi.acm.org/10.1145/1873951.1874249

[24] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE*, vol. 10, no. 3, pp. 1–21, 03 2015. [Online]. Available: https://doi.org/10.1371/journal.pone.0118432