31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

A Data Mining Algorithmic Approach for Processing Wireless Capsule Endoscopy Data Sets

Alexandros Karargyris and **Nikolaos Bourbakis**
College of Engineering, Assistive Technologies Research Center, Wright State University
Dayton, OH, 45435, USA
http://www.cs.wright.edu/atrc

Abstract: ***Wireless Capsule Endoscopy (WCE) has been a breakthrough in recent medical technology. It is used to view the gastrointestinal tract and detect abnormalities such as bleeding, Crohn's disease, peptic ulcers, and colon cancer. In this paper data mining techniques are utilized to extract useful information from a dataset of abnormal regions and non-abnormal regions. More specifically, the dataset contains polyps regions, ulcers regions and healthy regions. A number of features (shape descriptors, texture descriptors and color information) has been extracted for these regions and using a data mining toolbox useful conclusions are given on various relationships between these regions.***

Keywords: *Wireless capsule Endoscopy Imaging, data mining, imaging, polyps, ulcer, shape, texture*

I. Introduction – Wireless Capsule Endoscopy

WCE is a miniature camera capsule that is swallowed by the patient. It takes around 50,000 screenshots of the patient's digestive tract which are wirelessly transmitted to a wearable receiver. The 8-hour video is uploaded to a workstation for further viewing. With this technology physicians are able to search for gastrointestinal diseases and abnormalities such as Crohn's disease, ulcers, blood-based abnormalities and polyps.

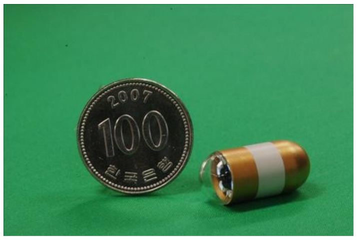

Figure 1. Miniature video capsule by Given Imaging[1]

Right now in the market there is a successful miniature capsule product, PillCam which is distributed by Given Imaging (fig. 1). So far various researches on WCE technology have been carried out that focus on three (3) goals: 1) automatic identification of abnormalities in WCE videos, 2) removal of redundant video information and 3) hardware improvement. In previous work [2][3] we presented methodologies that deal with the detection of ulcers and polyps in WCE videos. This paper extends that previous research and moreover, it tries to find useful relationships between the healthy tissue regions, polyps regions and ulcers regions. The rest of the paper gives a short description of polyps and ulcers, the extraction of the features, data mining techniques on the dataset, relationship graphs and finally conclusions.

Polyps are growing tissues inside the human body. Although polyps usually appear in the colon, stomach, and urinary bladder they may also appear in the small intestine. A picture of a colon polyp is given in figure 2, below.

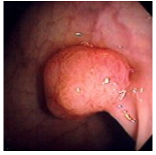

Figure 2. Colon polyp. Source: Stephen Holland, M.D., Naperville Gastroenterology, Naperville, IL, USA. Wikipedia.org

Polyps can either be benign or non-benign. Most polyps are not cancerous. However, since they can turn into cancer, physicians remove them and test them by performing biopsy [4]. Sometimes polyps can bleed causing anemia, while if they are bigger than a centimeter they have a greater cancer risk associated with them than polyps under a centimeter.

A peptic ulcer is defined as an area where tissue has been destroyed by gastric juices. Gastric juices are produced by the stomach and the intestine in order to digest the starch, fat, and protein in food. Since the intestine and the stomach also consist of proteins, they are protected by a) mucous layer, b) bicarbonate, which neutralizes acid and c) prostaglandins, which are hormones to boost bicarbonate and mucus production. In case these defense mechanisms are disturbed and acid and pepsin are allowed to attack the wall of the GI tract, ulcers may result. Although most peptic ulcers appear in the stomach (gastric ulcers) and the duodenum (duodenal ulcers) they may also appear in the small bowel. These small intestine ulcerations can be connected with various peptic diseases, such as lymphoma, carcinoma, and Crohn's disease. In some cases no specific causes can be found.

An ulcer can turn into three (3) types of complications: a) bleeding ulcer, b) perforated ulcer and c) narrowing ulcer. A bleeding ulcer occurs when the ulcer erodes one of the blood vessels. The mortality rate for bleeding peptic ulcers is about 10% [5]. Perforated ulcers appear as a hole in the wall and they are the typical case of peptic ulcers [6]. They can lead to intense abdominal pain. Narrowing ulcers (ulcerated strictures) cause stenosis (stricture) of the intestine and can lead to severe vomiting.

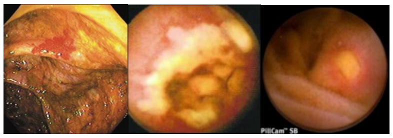

Figure 3. Three (3) types of ulcers. Left: bleeding ulcer, Middle: narrowing ulcer, Right: Perforated ulcer

III. Features

Our main goal in this research project was to have as many features as possible to describe the extracted regions: a) healthy tissue regions, b) polyps and c) ulcers.

It is ground truth that every natural object can be described by three (3) characteristics:

a) Geometry
b) Texture
c) Color

Keeping this in mind we tried to find features that offer as much information as possible for these characteristics. Concerning first category, geometry, there are many shape descriptors that can help measure regions. After careful investigation we used the following statistics: Area, Coordinates of region centroid, Major Axis Length, Minor Axis Length, Eccentricity, Orientation, Convex Hull, Equivalent diameter, Solidity, Extent, Perimeter. Since shapes in nature are not perfect circles but rather irregular major axis length, minor axis length and eccentricity describe a hypothetical ellipse that best fits the region. Convex hull or convexity is a feature that offers a degree of roughness of the region's periphery. Equivalent diameter computes the diameter of the circle that has the same area as the region. Solidity is a measure of the firmness/ compactness of the region. Extent gives a measure of how widely spread the region is inside a boundary box.

For the second category, texture, we used some of the texture descriptors proposed by Haralick. In his work [7] he describes fourteen (14) statistics that can be calculated from the co-occurrence matrix of the image. In our case, we utilized four (4) of these measurements: entropy, contrast, homogeneity and inverse moment.

For the last category we simply used the average values of red, green and blue components of the extracted regions. Additionally we calculated the average grayscale value of the regions.

Thus, we came up with a large multidimensional space (20-feature space) to describe the regions efficiently.

IV. Data Mining – Information Extraction

Data mining is a process that helps discover uncovering patterns, associations and structures inside data. With data mining we try to find correlations or patterns among features stored in databases.

Before we are able to perform data mining we need to create a dataset with the three (3) types of tissues. The feature extraction methods are described in previous publications ([2],[9]).Based on our methodologies, we extracted from various WCE videos (totally 3000 frames) forty (40) ulcer regions, forty (40) polyps regions and a hundred twenty (120) healthy tissue regions that were potentially either polyp candidates or ulcer candidates. Thus, the total measurements were four thousand (4000). The acquisition of real data is hard since it takes long time to collect sufficient amount of abnormalities cases. Even though the amount of 4000 measurements could be arguable it is sufficient to give us useful information and indications on the relations of abnormalities. A simple flowchart that demonstrates our data mining is given below in figure 4.

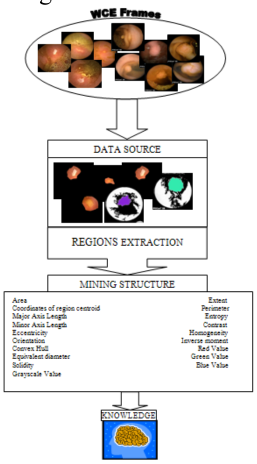

Figure 4. Data mining flowchart

As data mining software we used WEKA 3-6-0 developed by the University of Waikato [8]. WEKA toolbox offers various options: pre-processing, classification, clustering, features associating and of course visualization. Using this toolbox we were able to perform experiments on our dataset easily and efficiently.

V. Experiments

We performed two (2) types of experiments with our dataset. In the first experiment we used the whole dataset to try to find correlations between the three (3) types of tissues. In the second experiment we used only the two (2) abnormal tissues. Both experiments had two phases: i) classification and ii) associations. In classification step we evaluated various classifiers to find those which could classify the tissues types more accurately. This was part of an ongoing effort to increase the accuracy of detection of polyps and ulcers. In associations step we tried to find hidden correlations between tissues.

5.1 Polyps vs Ulcers vs Normal Tissues

Weka software offers a wide range of classifiers to use. In table 1 we are presenting the top ten (10) classifiers that performed well. We decided to use the default parameters of each classifier. As one can see from that table, nearest neighbor classifier (fig. 5) behaved the best for all categories of tissues. IB1 classifier worked perfectly only for ulcers whereas Simple Logistic classifier worked best for healthy tissues. It can also be seen that polyps' detection didn't reach sensitivity above 70%. The reason for this is that polyps have shape, texture and color very similar to normal tissues, making it hard to distinguish them. On the other hand, ulcers reached a great level of sensitivity and specificity which can be explained mainly due to the difference in texture and color from polyps and healthy tissues.

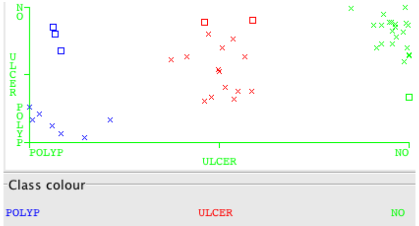

Figure 5. Nearest Neighbor with generalization rule is the toppest classifier in our dataset. Xs are instances that were assigned to the correct category whereas □ are instances that were assigned to incorrect category.

Table 1. Sensitivity, (1-specificity) and precision for each tissue

Polyps

	True Positive	False Positive	Precision
Nearest Neighbor with generalization rule	0.7	0.0	1.0
Logistic model tree	0.6	0.051	0.75
Naive Bayes/Decision-Tree	0.5	0.077	0.625
Fuzzy Lattice Reasoning (FLR) classifier	0.4	0.0	1
Voting feature intervals classifier	0.7	0.179	0.5
Classification via Regression	0.5	0.051	0.714
Instance-based learning (IB1) classifier using 1 nearest neighbor(s) for classification	0.6	0.051	0.75
KStar Beta Verion	0.7	0.256	0.412
NN Multilayer Perceptron	0.6	0.026	0.857
Simple Logistic classifier	0.6	0.051	0.75

Ulcers

	True Positive	False Positive	Precision
Nearest Neighbor with generalization rule	0.882	0.031	0.938
Logistic model tree	0.941	0.031	0.941
Naive Bayes/Decision-Tree	0.941	0.031	0.941
Fuzzy Lattice Reasoning (FLR) classifier	0.882	0.0	1
Voting feature intervals classifier	0.941	0.031	0.941
Classification via Regression	0.824	0.0	1
Instance-based learning (IB1) classifier using 1 nearest neighbor(s) for classification	1	0.0	1
KStar Beta Verion	0.588	0	1
NN Multilayer Perceptron	1	0	1
Simple Logistic classifier	0.941	0.031	0.941

Healthy tissues

	True Positive	False Positive	Precision
Nearest Neighbor with generalization rule	0.955	0.185	0.808
Logistic model tree	0.864	0.185	0.792
Naive Bayes/Decision-Tree	0.864	0.185	0.792
Fuzzy Lattice Reasoning (FLR) classifier	1	0.296	0.733
Voting feature intervals classifier	0.636	0.148	0.778
Classification via Regression	0.909	0.296	0.714
Instance-based learning (IB1) classifier using 1 nearest neighbor(s) for classification	0.909	0.148	0.833
KStar Beta Verion	0.636	0.296	0.636
NN Multilayer Perceptron	0.955	0.148	0.84
Simple Logistic classifier	0.864	0.185	0.892

Furthermore, we tried to find various associations between the features of the patterns. Weka offers the ability to view 2D graphs between two (2) features for all classes and also identify any associations.

Graphs proved very helpful since they were able to display correlations that could help distinguish the classes. We knew beforehand that eccentricity plays a major role in detection of polyps but combined with contrast feature it proved to increase the sensitivity of this detection (fig.6).

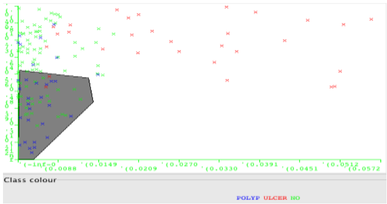

Figure 6. Contrast vs. Eccentricity (shadow area shows the majority of polyps)

Elaborating on why the contrast feature helped above in fig. 6 it can be said that this is reasonable since contrast makes an object distinguishable from other objects (background). We moved this thinking a step further and we had a discussion with our associate physician Dr. Marios Pouagere. From this discussion we understood that besides the shape of the polyps (eccentricity), physicians, without realizing it but due to their experience, use the concept of contrast to recognize polyps and extract them from the rest of the WCE frame.

Encouraged by this discovery we further investigated other features. In case of the extent feature we noticed that polyp instances are more concentrated in a certain area than the rest of the two classes (fig. 7).

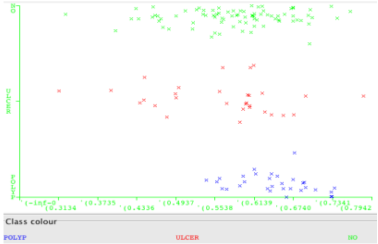

Figure 7. Classes vs. Extent

This motivated us to search for other features that could help distinguish polyps. In our case extent vs. convexity increased the sensitivity of the detection. This can be explained to the fact that having applied our algorithm the extracted polyps are crisper regions with distinct borders. Since these regions are created naturally it is really difficult that they will be round shaped. On the contrary, they have protrusions and intrusions creating concave areas and thus higher degree of convexity (fig. 10). Figure 8 shows exactly that polyps have higher value in convexity.

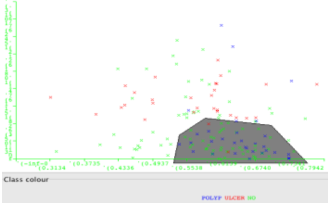

Figure 8. Extent vs. Convexity

Another really useful conclusion was the coordinates of the regions of each category. Most of polyps and ulcer were concentrated near the center of the frame whereas the healthy regions were scattered all over the frame (fig. 9). This shows that it is wiser for our algorithm to look inside a perimeter of the frame than scan for the whole frame for abnormalities.

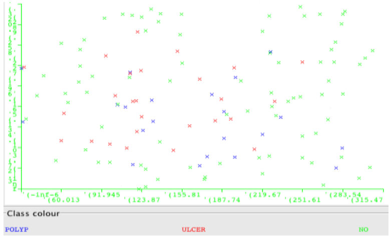

Figure 9. X coordinates and Y coordinates of classes

Weka offers an associator tool to identify associations between features. In our case an interesting association came up: if the instance belongs to a normal class then the inverse moment value is below 39.606 with confidence 94%. Another association is that if convexity is below 5013.9 then the area is smaller than 4571 pixels with confidence of 93%. This means that medium sized-regions have quite relatively large convexity showing the roughness of the perimeter of the region (fig. 10).

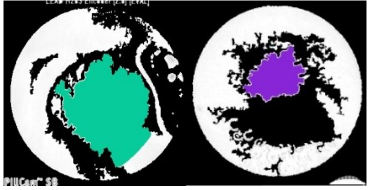

Figure 10. Typical extracted polyps. Roughness of the perimeter is obvious

5.2 POLYPS vs. ULCERS

We decided to examine the two abnormalities, polyps and ulcers, and search for any correlations and associations.

Polyps and ulcers have more differences than similarities, which is normal. Our assumption was that the polyps are more round shaped regions than ulcers. However, going through the data set we found out that ulcers were closer to circle shape than polyps (fig 11).

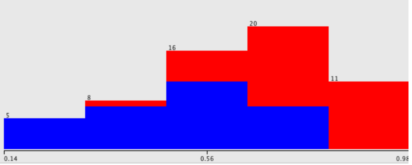

Figure 11. Eccentricity feature for polyps (blue) and ulcers (red)

Surprisingly enough, ulcers proved to cover averagely a bigger area than polyps (fig. 12). After all, that is well understood since ulcers consist of two areas (inner and outer areas, see figure 3).

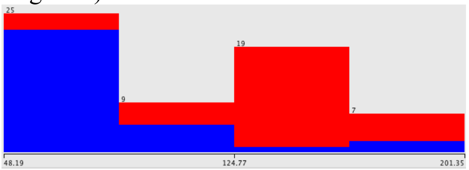

Figure 12. Equivalent diameter feature for polyps (blue) and ulcers (red)

Furthermore, ulcers have higher value of entropy than polyps. This can be explained from the fact that ulcers have intense texture whereas polyps are more homogenous.

Comparing two (2) features at a time we figured out that polyps have contrast and solidity values in narrow ranges, whereas ulcers have their contrast and solidity values widely spread. This makes us believe that physicians might be able to distinguish ulcers using mainly color rather than shape and texture, whereas in case of polyps it comes down to shape and contrast. Our assumption seems to be backed by the fact that ulcers have a dominant average color value whereas polyps do not (fig. 13).

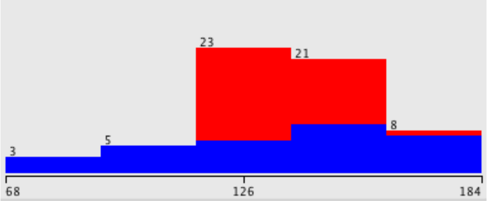

Figure 13. Average color value

VI. Conclusions

In this paper we primarily had two goals: 1) create a big feature space and 2) evaluate these features so that to come up with meaningful conclusions about abnormal and normal regions. It is surprising to see that getting into the data sets and looking at them from every perspective led us to see some new correlations and confirm some assumptions.

Our experiments focused mainly on all three (3) classes rather than just the two (2) of them –polyps and ulcers. The reason was that it was more challenging for us to try to find ways to distinguish the three (3) classes. Besides that, the evaluation of the classifiers in 5.1 is going to help us implement the best classifier into our software for the detection of polyps.

Finally, our data mining helped us comprehend scientifically how physicians perceive certain patterns. Before this we lacked this knowledge and we were merely making assumptions on how medical personnel examine the WCE videos. This is probably one of the most important contributions of our data mining: try to understand how humans perceive certain patterns and thus, try to mimic their behavior and perception to create smart algorithms.

In the future we plan to create an even larger dataset including other pathological patterns such as blood-based abnormalities. Additionally we want to expand the feature space to include time (frame number in the case of WCE videos) to see if there are conditional existences of abnormalities.

References

[1] Source: Yonhapnews.co.kr

[2] Karargyris Alexandros; Bourbakis, Nikolaos , "A Video-frame based Registration using Segmentation and Graph Connectivity for Wireless Capsule Endoscopy" , Life Science Systems & Application Workshop (LiSSA '09) IEEE / NIH

[3] Karargyris Alexandros; Bourbakis, Nikolaos , "Identification of polyps in Wireless Capsule Endoscopy videos using Log Gabor filters", Life Science Systems & Application Workshop (LiSSA '09) IEEE / NIH

[4] What I need to know about Colon Polyps? National Institute of Diabetes and Digestive and Kidney Diseases, National Institutes of Health. http://digestive.niddk.nih.gov/ddiseases/pubs/colonpolyps_ez/ - Viewed in November 2008.

[5] Well-Connected Reports, Harvard Medical School, Peptic Ulcers, September 2001

[6] Umaprasanna S. Karnam, Charles M. Rosen, Jeffrey B. Raskin - Small bowel ulcers, Current Treatment Options in Gastroenterology Journal, Vol. 4, Issue 1, pp 15-21

[7] Robert M. Haralick, ``Statistical and structural approaches to texture,'' Proc. IEEE, vol. 67, no. 5, pp. 786-804, 1979.

[8] Weka 3: Data Mining Software, University of Waikato , http://www.cs.waikato.ac.nz/ml/weka/

[9] Karargyris Alexandros; Bourbakis, Nikolaos , "Identification of ulcers in Wireless Capsule Endoscopy" IEEE International Symposium on Biomedical Imaging 2009