LEARNING DISEASE SEVERITY FOR CAPSULE ENDOSCOPY IMAGES

R. Kumar[*], *P. Rajan, S. Bejakovic, S. Seshamani, G. Mullin*[†], *MD, T. Dassopoulos*[‡], *MD, G. Hager*[§]

Department of Computer Science, Johns Hopkins University, Baltimore, MD
Johns Hopkins Hospital, Baltimore, MD
Washington University, School of Medicine St. Louis, MO

ABSTRACT

Wireless capsule endoscopy (CE) is increasing being used to assess several gastrointestinal(GI) diseases and disorders. Current clinical methods are based on subjective evaluation of images. In this paper, we develop a method for ranking lesions appearing in CE images. This ranking is based on pairwise comparisons among representative images supplied by an expert. With such sparse pairwise rank information for a small number of images, we investigate methods for creating and evaluating global ranking functions. In experiments with CE images, we train statistical classifiers using color and edge feature descriptors extracted from manually annotated regions of interest. Experiments on a data set using Crohn's disease lesions for lesion severity are presented with the developed ranking functions achieve high accuracy rates.

***Index Terms*—** Capsule Endoscopy, Statistical Classification, Disease Severity, Ordinal Regression

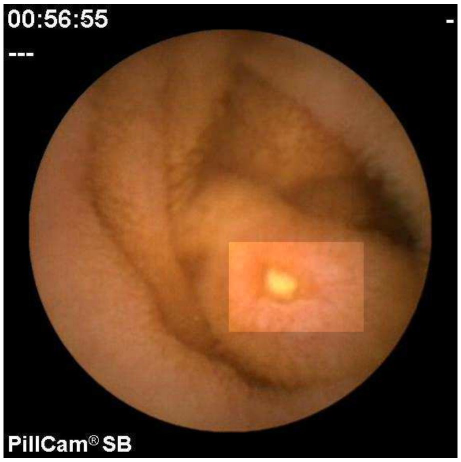

Fig. 1. CE image of a Crohn's disease lesion with highlighted ROI showing the lesion and surrounding inflammation.

1. INTRODUCTION

Wireless capsule endoscopy (CE) [1] is increasingly being used to diagnose small bowel conditions such as obscure gastrointestinal bleeding, celiac disease, and Crohn's disease. The disposable capsule system (Given Imaging Inc, or Olympus Medical Systems), not much larger than a common drug capsule, consists of a small color camera, LEDs and electronics for illumination and wireless communication, and the battery. The capsule is swallowed and moved by peristalsis along the small intestine. The device typically transmits approximately 50,000 images at a rate of 2 (typically 576x576, color) images per second for up to 8 hours to a wireless receiving device worn on the body, limited only by battery life.

The archived images are later analyzed by a clinician in a potentially time consuming process. The typical study reading time is reported to be in hours [1, 2]. In addition to being a tedious, detection rates may also vary among clinicians, especially for early stage disease. As a result, consistent assessment of disease severity in CE images poses several challenges. It requires consistently detecting and assessing lesions in video streams. It also requires recognizing and accounting for redudant views of the same lesion. Therefore, it is important to devise methods for efficient, consistent lesion assessment.

This paper explores methods for automating assessment of lesion severity in capsule endoscopy. Classification and ranking, formulated as problems of learning a map from a set of feature to a discrete set of labels, have been applied widely in computer vision applications for face detection [3], object recognition [4], and scene classification [5]. Alternatively, ranking can be viewed as a *regression* problem to find a ranking function between a set of input features and a continuous range of ranks or assessment. This form has gained recent interest in many areas such as learning preferences for movies (http://www.netflixprize.com), or learning ranking functions for web pages (e.g. google page rank).

Learning ranking functions requires manually assigning a consistent ranking scale to a set of training data. Although the scale could be arbitrary, what is of interest is the consis-

[*] email:rajesh@jhu.edu

[†] Dr. Mullin is affiliated with the Johns Hopkins Hospital

[‡] Dr. Dassopoulos is with Washington Univestiy, St.Louis

[§] Supported in part by National Institutes of Health with Grant 5R21EB008227-02 and Johns Hopkins University internal funds

tent ordering of the sequence of images; a numerical scale is only one of the possible means of representing this ordering. Ordinal regression tries to learn a ranking function from a training set of partial order relationships. The learned global ranking function then seeks to respect these partial orderings while assigning a fixed rank score to each individual image or object. Both Machine learning [6, 7] and content based information retrieval [8] have sought to obtain mapping functions assigning preference or ranking scores. In our work, we use selective sampling techniques and SVMs with user provided sparse partial ordering in combination with image feature vectors automatically generated from a training set of images.

2. METHODS

Consider a vector of training images $\mathcal{I} = [I_1, I_2 ... I_n]$. A subset of $\mathcal{I}$ have an associated preference relationship $\prec$. Let

$$\mathcal{P} = \{(x, y) \mid I_x \prec I_y\}.$$

Let $\bar{P}$ denote the transitive closure of $\mathcal{P}$. We require that $(x, x) \notin \bar{P}$, thus disallowing inconsistent preferences. Our goal is to to compute a real-valued ranking function R such that

$$I_x \prec I_y \in P \implies R(I_x) < R(I_y)$$

For the rest of this discusion, "rank" will refer to a real-valued measure on a linear scale, and "preference" will denote a comparison among objects. We note that, given a numerical ranking on n items, it is straightforward to generate $O(n^2)$ preference relationships. Likewise, given a categorization of n items into one of m bins on a scale (e.g. mild, moderate, or severe lesion), it is again possible to generate $O(n^2)$ preferences. Thus, this formulation subsumes both scale classification and numerical regression.

Our estimate of the ranking function is based on empirical statistics of the training set. The key idea is to note that a preference pair $\langle x, y \rangle \in \bar{P}$ can be thought of as a *pair of* training examples for a binary classifier. Let us define

$$B(p) = \begin{cases} 0 & p \in \bar{P} \\ 1 & \text{otherwise} \end{cases} \qquad (1)$$

We train a classifier C such that for any $p \in \bar{P}$

1. $C(I_x, I_y) = B(\langle x, y \rangle)$

2. $C(I_y, I_x) = 1 - B(\langle x, y \rangle)$

Given such a classifier, a continuous valued ranking can be easily produced as

$$R(I) = \sum_{i=1}^{n} C(I_i, I)/n \qquad (2)$$

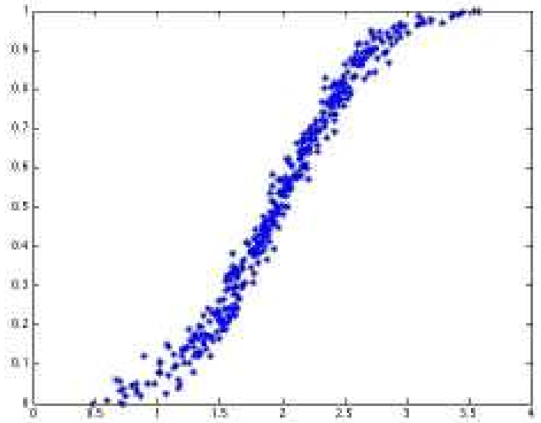

Fig. 2. Estimated ranks vs. feature vector sum ($\sum f$) for simulated data.

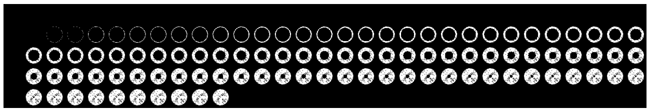

Fig. 3. Disc images sorted (left to right) by estimated ranks.

That is, R is the fraction of values of the training set that are "below" I based on the classifier. Thus, R is also the empirical order statistic of I relative to the training set. The formulation above can be paired with nearly any binary classification algorithm.

Here, we will make use of SVMs in combination with feature vectors extracted from the CE images. We assume that an I_x is represented by a feature vector f_x. As training examples require pairs of images, let $f_{k,j}$ represent the vector concatenation of f_k and f_j. The training set thus consists of the set

$$\mathcal{T} = \{< f_{k,j}, 0 >, < f_{j,k}, 1 > |(k, j) \in \bar{P}\}$$

The result of performing training on $\mathcal{T}$ is a classifier which, given a pair of images, will determine their relative order. Give this framework, we investigate preference relationships needed for training, and image features that produce robust feature vectors.

By way of illustration, consider random vectors in R^4 with the following preference rule: $f_1 \prec f_2$ if and only if $\sum f_1 < \sum f_2$. The ranking function $\mathcal{R}$ obtained from an SVM classifier trained on 200 samples is plotted versus $\sum f$ in Figure 2. The training set included all available feature vectors, and achieved a 0% misclassification rate.

As a second example, consider a set of 100 synthetic images of disks (Figure 3) of varying thickness. Each image is 131x131 and grayscale, with the disc representing the only non-zero pixels, consecutive images differing by 0.5 pixels in disc thickness. For images I_i and I_j, the underlying ranking function is $thickness(i) < thickness(j) \equiv i \prec j$. Using a 10 bin intensity histograms as the feature vector, a SVM classifier using radial basis functions produces a ranking function $\mathcal{R}$ that correctly orders (0 % misclassification) the discs (Figure 3) using only $O(n)$ pairwise relationships.

3. EXPERIMENTS

We have an ongoing Johns Hopkins Medical Institutions (JHMI) Institutional Review Board (IRB) approved protocol for collecting anonymized capsule endoscopy studies. These studies are anonymized and reviewed by Dr. Dassopoulos for assessment of Crohn's disease lesions. During this review, lesions as well as data for other classes for interest are selected and assigned a global ranking (mild, moderate, or severe) based upon the size, and severity of lesion and any surrounding inflammation. Lesions are ranked into three categories: mild, moderate or severe disease.

A region of interest (ROI) is also manually computed. Figure 1 shows a typical Crohn's disease lesion. As a lesion may appear in several images, data representing 50 seconds of recording time around the selected image frame is also reviewed, annotated, and exported as a sequence. In addition, a number of extra image sequences not containing lesions are exported as background data for training of statistical methods.

We use the global lesion ranking to generate the required preference relationships $\prec$. Over 188,000 pairwise relationships are possible over our selected dataset of 600 lesion image frames that have been assigned a global ranking of mild, moderate or severe by the clinician, assuming $mild < moderate < severe$. We utilize a small number to initiate training, and an additional number to iterate for improvement of the ranking function. Previous work on machine learning has generally made use of some combination of color and texture features. SIFT [9] is not very suitable for our wireless endoscopy images, due to lack of sufficient number of SIFT features in these images. A variety of feature vectors including edge, color, and texture features [10], MPEG-7 visual descriptors [11], and hue, saturation and intensity features [12] have been published specifically for analysis of wireless capsule endoscopy images. In prior work, we [13] have also explored feature vector representations based on color, edge and texture information for our lesion data. For these experiments, simple 10 bin normalized hue and saturation histograms provide adequate feature vectors.

In the experiments below, we explore the improvement of accuracy of the ranking function with increasing number of pairwise preferences. These experiments were performed using the SVM library in MATLAB's bioinformatics toolbox, on a Solaris 5.10 cluster of 64 processors (each running at 1167 MHz).

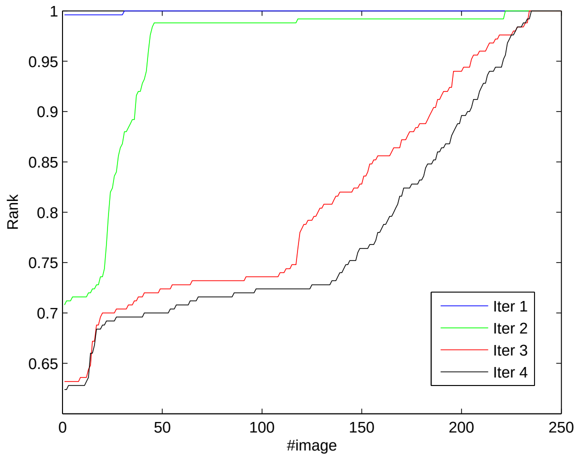

Fig. 4. Iterations of the estimated rank function for 250 lesion images.

Metric	Iter. 2	Iter. 3	Iter. 4
Mean	0.1133	0.0182	0.0024
Std. Dev	0.2055	0.0915	0.0106

Metric	Iter. 1	Iter. 2	Iter. 3	Iter. 4
Training size	100	1100	1972	2116
mismatches	1286	436	77	3

Table 1. Changes in rank (top) and mismatches (bottom) over iterations for 100 lesion images.

On $n = 100$ images, starting with only $O(n)$ training relationships, and SVM classifier using radial basis functions as before, we obtain only $O(n^2)$ mismatches using the generated ranking function $\mathcal{R}$ after the first iteration. A mismatch is any pair of images where $R(I_x) < or > R(I_y)$ and $I_x > or < I_y$ The number of mismatches drops exponentially over 4 iterations where the training set is increased by $m = max(1000, mismatches)$ pairwise relationships.

Figure 6 shows the resulting ranked images, and Table 1 show changes in ranks for images, and number of mismatches during each iteration. Both the mean and standard deviation of rank change for individual images decreases monotonously over successive iterations. Table 1 also shows the decreasing number of mismatches over successive iterations. Figure 4 shows the sorted ranking function for 250 images over iterations. The ranking function converges after just a few iterations, with the changes in rank becoming smaller closer to the convergence. Figure 5 show the decrease in mismatches for 100 and 250 images over 4 iterations. Finally, Figure 7 contains similarly ranked 500 lesion images.

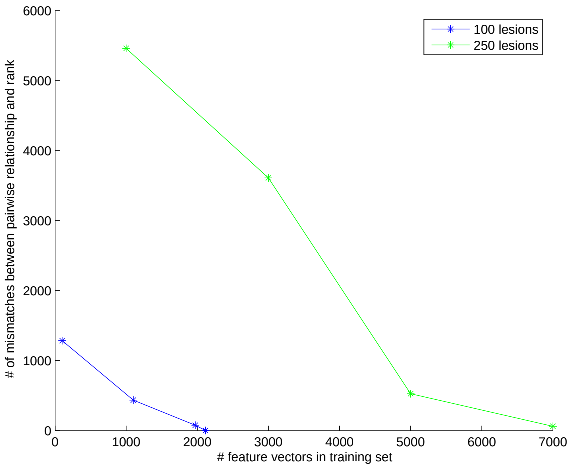

Fig. 5. Number of iterations compared to the number of preference relationships not respected by the estimated ranks for 100, and 250 lesion images.

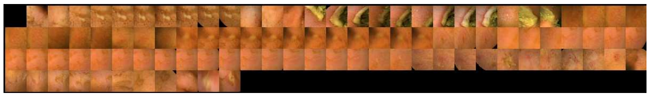

Fig. 6. A montage of 100 ranked lesion images.

4. CONCLUSIONS

We have experimented with a framework using only O(n) relationships to establish useful ordinal assessment of feature vectors automatically extracted from a set of medical images. While this work only establishes global assessment rankings, this can be extended to obtain finer relationships, for example, size, shape, and depth properties of lesions and surrounding inflammation for our lesions data set. Suitable compositions of such rankings may provide assessments with very high correlation with manual assessments.

The lesion rankings used here were generated by a single expert clinical user. In our continuing work, we are now performing blind review of these images to improve the consistency of manual annotation. The ROIs used were also manually annotated, and while this is not a significant problem for the small number of preference relationships required, we will also explore suitable methods for automatic segmentation of ROIs in the future.

5. REFERENCES

[1] G. Iddan, G. Meron, A. Glukhovsky, and P. Swain, “Wireless capsule endoscopy,” *Nature*, vol. 405, no. 6785, pp. 417, 2000.

[2] B.S. Lewis, “Expanding role of capsule endoscopy in inflammatory bowel disease,” *World J Gastroenterol*, vol. 14, no. 26, pp. 4137–4141, 2008.

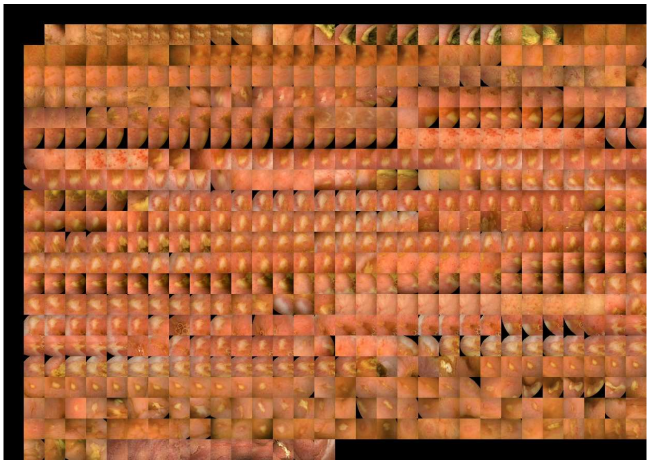

Fig. 7. A montage of 500 lesion images. Note that multiple views of the same lesion are ranked close to each other in sorted montages.

[3] P. Voila and M. Jones, “Robust real-time face diction [J],” *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[4] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer, “Generic Object Recognition with Boosting,” *IEEE PAMI*, pp. 416–431, 2006.

[5] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, “Learning object categories from google’s image search,” in *Proc. ICCV*, 2005, pp. 1816–1823.

[6] J. Furnkranz and E. Hullermeier, “Pairwise Preference Learning and Ranking,” *Lec. Notes in Comp. Sc.*, pp. 145–156, 2003.

[7] R. Herbrich, T. Graepel, and K. Obermayer, *Regression Models for Ordinal Data: A Machine Learning Approach*, Technische Universität Berlin, 1999.

[8] S. Tong and E. Chang, “Support vector machine active learning for image retrieval,” in *Proc. of 9th ACM Int. conf. on Multimedia*. ACM New York, NY, USA, 2001, pp. 107–118.

[9] D.G. Lowe, “Object recognition from local scale-invariant features,” in *Proc. ICCV*. Kerkyra, Greece, 1999, vol. 2, pp. 1150–1157.

[10] Y. Liu, D. Zhang, G. Lu, and W.Y. Ma, “A survey of content-based image retrieval with high-level semantics,” *Pattern Recognition*, vol. 40, no. 1, pp. 262–282, 2007.

[11] M. Coimbra, P. Campos, and JPS Cunha, “Topographic Segmentation and Transit Time Estimation for Endoscopic Capsule Exams,” in *Proc. ICASSP*, 2006, vol. 2.

[12] Jeongkyu Lee, JungHwan Oh, Subodh Kumar Shah, Xiaohui Yuan, and Shou Jiang Tang, “Automatic classification of digestive organs in wireless capsule endoscopy videos,” in *SAC07*, 2007.

[13] Srdan Bejakovic, Rajesh Kumar, Themistocles Dassopoulos, Gerard Mullin, and Gregory Hager, “Analysis of crohns disease lesions in capsule endoscopy images,” in *IEEE ICRA*, 2009(accepted).