Learning from Partially Annotated OPT Images by Contextual Relevance Ranking

Wenqi Li[1], Jianguo Zhang[1], Wei-Shi Zheng[2],
Maria Coats[3], Frank A. Carey[3], and Stephen J. McKenna[1]

[1] CVIP, School of Computing, University of Dundee, Dundee, UK
[2] School of Information Science and Technology, Sun Yat-sen University, China
[3] Ninewells Hospital & Medical School, University of Dundee, UK

Abstract. Annotations delineating regions of interest can provide valuable information for training medical image classification and segmentation methods. However the process of obtaining annotations is tedious and time-consuming, especially for high-resolution volumetric images. In this paper we present a novel learning framework to reduce the requirement of manual annotations while achieving competitive classification performance. The approach is evaluated on a dataset with 59 3D optical projection tomography images of colorectal polyps. The results show that the proposed method can robustly infer patterns from partially annotated images with low computational cost.

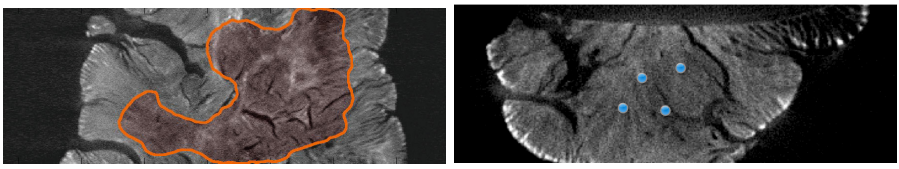
1 Introduction

Optical Projection Tomography (OPT) microscopy is a relatively new 3D imaging modality [7]. It has an effective resolution of 5μm to 10μm and is ideally suited for specimens between 0.5 mm and 10 mm. Recently OPT has been used to image colorectal polyps. The analysis of these images is currently performed visually and the classification of polyps exhibits variability depending on the experience and awareness of the experts [4]. We investigate automated analysis of polyp regions to assist pathologists in colorectal cancer diagnosis.

To model the underlying patterns of image regions, accurate annotations are desirable. However the volumetric images of polyps are large (1024^3 voxels); while high resolution brings us considerable detail, difficulty arises in obtaining annotations. In our dataset, a polyp typically extends across $700 \sim 800$ slices and about 0.5 billion voxels in total. Fully delineating 3D regions slice by slice is tedious and time-consuming.

In this paper we investigate an alternative approach based on partial, sparse, incomplete annotations. We propose a learning framework for partially annotated OPT images, for the task of classifying dysplastic changes in colorectal polyps. More specifically, the objective in this paper is to discriminate between image patches that contain low-grade dysplasia (LGD) and image patches that contain invasive cancer. This is a first step towards the goal of automated polyp analysis.

K. Mori et al. (Eds.): MICCAI 2013, Part III, LNCS 8151, pp. 429–436, 2013.



(a) (b)

Fig. 1. OPT colorectal polyp images with (a) region annotation and (b) partial annotations

Different forms of partial annotation can be appropriate for different image modalities and applications. In this paper, we consider partial annotations consisting of just one click or a few clicks in the 3D polyp region of interest (as shown in Fig. 1(b)) as an alternative to the stronger annotation shown in Fig. 1(a). The annotation effort required is quite different. Our goal is to reduce the annotation efforts while achieving good classification performance. In addition, learning should scale well making it suitable for high-resolution volumetric images.

In [4] local features for patch and region classification of OPT images were compared. Here we focus on the model learning aspect of the task. Our method falls into the broad category of weakly supervised classification. At one extreme of this category, annotation is performed only at the image level in which case Multiple Instance Learning (MIL) has been adopted. In MIL, a sample is classified as positive if at least one of the instances is classified as positive. Dundar et al. [2] proposed a large margin based approach for pathology slides. It shared some similarity to this work however the prediction was at image level. In [8] MIL was adapted to classify and segment histopathology images. Doyle et al. [1] applied active learning to detect cancer regions with histopathology annotations. Our approach is to leverage spatial annotation but to keep annotation simple, sparse and thus fast to perform.

2 Methods

In supervised classification settings, locations outside annotated regions are usually ignored during training because the corresponding class labels are considered unknown. However, for images annotated with a partial annotation protocol, the annotations carry information about the class membership at unannotated locations. We refer to 3D windows as *patches*. Patches in the training set at locations with annotated (known) class labels are referred to as *reference patches*. Patches near to them (in terms of displacement or distance in feature space) are referred to as *candidate patches*. In this paper, we consider an extreme form of partial annotation consisting of point locations obtained via mouse clicks. We introduce our definition of contextual relevance, based on which we then propose a ranking model for classification.

2.1 Labeling Patches' Confidence

First we assign confidence labels to candidate patches. Consider a reference patch $\mathbf{S}_r$ sampled at an annotated location $\mathbf{z}_r$ labeled as $y_r \in \{1, -1\}$. The patch $\mathbf{S}_k$ sampled at location $\mathbf{z}_k$ will have a lower confidence label y_k. y_k can be set to:

$$y_k = a(\mathbf{S}_k, \mathbf{S}_r)\, y_r, \tag{1}$$

where $a(\cdot\,, \cdot) \in (0, 1]$ is a measurement of affinity between two image patches. The absolute value of y_k can be viewed as a confidence measurement.

As patches sampled at locations near to each other usually belong to the same class, the reference patch of S_k can be set as the nearest annotated patch S_r. Affinity $a(\cdot\,, \cdot)$ is set as a Gaussian function with regard to spatial displacement of S_k and S_r in the image and a scaling parameter σ, i.e.:

$$a(\mathbf{S}_k, \mathbf{S}_r) = \exp(-\frac{\|\mathbf{z}_k - \mathbf{z}_r\|^2}{\sigma^2}). \tag{2}$$

Another way to define $a(\cdot\,, \cdot)$ is to consider similarity in image feature space. Assuming feature $\mathbf{x}_r$ is extracted from reference patch $\mathbf{S}_r$ and $\mathbf{x}_k$ from $\mathbf{S}_k$, we can alternatively define the function as:

$$a(\mathbf{S}_k, \mathbf{S}_r) = \exp\,(-\frac{\|\mathbf{x}_k - \mathbf{x}_r\|^2}{\delta^2}). \tag{3}$$

Note that $a(\cdot\,, \cdot)$ can be extended to use multiple reference patches. Here we assign only one (the nearest) reference patch for each candidate patch.

2.2 Contextual Relevance Ranking Model

Let $\mathbf{x}_i \in \mathbb{R}^d, i = 1, \cdots, N$ denote a feature vector extracted from an image patch indexed by i. We assign the label $y_i \in \{1, -1\}$ if the ith feature vector is from a reference patch; otherwise we set y_i according to formula (1). We form the ranking model by optimizing a regularized margin-based problem:

$$\min_{\mathbf{w},\, b} \quad \frac{\lambda}{2}||\mathbf{w}||^2, \tag{4}$$

$$\text{s.t.} \quad \frac{1}{y_i}(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \quad \forall \mathbf{x}_i \in \mathbf{X}, \tag{5}$$

$$\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j \geq R_{ij}, \quad \forall \mathbf{x}_i \in \mathbf{X}^+, \mathbf{x}_j \in \mathbf{X}^-, \tag{6}$$

where $\mathbf{X}^+ = \{\mathbf{x}_k : 0 < y_k < 1\}$ and $\mathbf{X}^- = \{\mathbf{x}_k : -1 < y_k < 0\}$. R_{ij} is the pairwise contextual relevance of two patches $\mathbf{S}_i$ and $\mathbf{S}_j$:

$$R_{ij} = \frac{a(\mathbf{S}_i, \mathbf{S}_{ir})a(\mathbf{S}_j, \mathbf{S}_{jr})}{a(\mathbf{S}_i, \mathbf{S}_{ir}) + a(\mathbf{S}_j, \mathbf{S}_{jr})}, \tag{7}$$

where the patches $\mathbf{S}_{ir}$ and $\mathbf{S}_{jr}$ are the reference patches of $\mathbf{S}_i$ and $\mathbf{S}_j$ respectively.

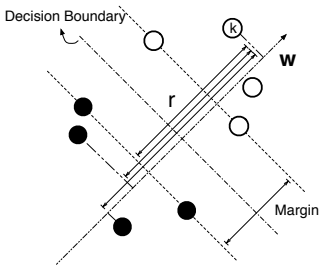


Fig. 2. Geometric interpretation of contextual relevance ranking model. **w** is the weight vector; point k is a feature vector extracted from a reference patch. r refers to the differences of ranking score between data point k and points in the other class (projected along the direction of **w**). Constraints in formula (5) were designed for minimizing classification error; constraints in formula (6) were designed for optimizing ranking difference r.

Constraints (5) are for all feature vectors in the training set. Note that in (5) features from candidate patches y_k are loosely constrained compared to their reference patches $y_r \in \{-1, 1\}$ because $|y_k| \in (0, 1)$. Constraints (6) are rankings of a pair of patches from two images with regard to their contextual relevance. We argue that patches sampled nearer to annotated locations (in terms of image location or location in feature space) should be classified with a larger score, i.e., further away from decision boundaries. The constraints keep the projected distance between any data point with high magnitude in y and the data in the opposite class large. In the case that pair ($\mathbf{x}_i$, $\mathbf{x}_j$) is labeled with certainty, i.e., $y_i = \pm 1$ and $y_j = \pm 1$, the pairwise constraint (6) vanishes due to constraint (5). Fig. 2 illustrates the geometric interpretation of this model.

Given a fixed training set, the optimization problem can be transformed into dual form of **w** by constructing a new feature set with $\frac{(\mathbf{x}_i - \mathbf{x}_j)}{R_{ij}}$. Then this can be solved by any SVM dual form solver, e.g. LIBSVM, SVM^{light}. However, this method is very slow and constructing feature set $\frac{(\mathbf{x}_i - \mathbf{x}_j)}{R_{ij}}$ across all the pairwise constraints is infeasible for our problem because the set of candidate patches is large. Here we tackle the primal form directly with a recently proposed efficient stochastic gradient method, SAG [6]. This method enables us to learn features online and with minimal storage cost.

To solve the optimization problem we minimize function (4) while controlling constraint violations in (5) and (6). Combining them together the risk function $J(\mathbf{w}, b)$ on training features $\{\mathbf{x}_i\}_{i=1}^N$ and labels $\{y_i\}_{i=1}^N$ can be written as:

$$J(\mathbf{w}, b) = \frac{\lambda}{2}||\mathbf{w}||^2 + \frac{1}{N^+N^-} \sum_{\mathbf{x}_i \in \mathbf{X}^+} \sum_{\mathbf{x}_j \in \mathbf{X}^-} (f_i + f_j + C g_{ij}); \tag{8}$$

$$\text{where: } f_i = f(\frac{1}{y_i}(\mathbf{w}^T\mathbf{x}_i + b)), \quad f_j = f(\frac{1}{y_j}(\mathbf{w}^T\mathbf{x}_j + b)), \tag{9}$$

$$g_{ij} = g(\mathbf{w}^T\mathbf{x}_i - \mathbf{w}^T\mathbf{x}_j - R_{ij}). \tag{10}$$

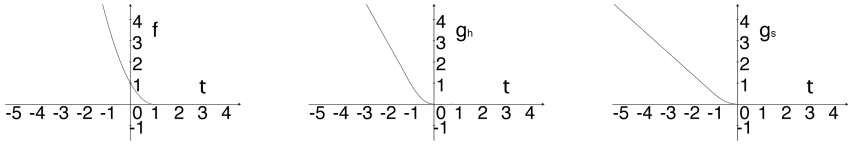


Fig. 3. Demonstration of loss functions (left to right): (1) squared hinge loss $f(t)$, (2) Huber loss $g_h(t)$, (3) smoothed hinge loss $g_s(t)$

The loss term $f(\cdot)$ corresponding to constraints (5) is squared hinge loss:

$$f(t) = \max(0,\, 1-t)^2; \tag{11}$$

the loss term $g(\cdot)$ corresponding to constraints (6) can be a Huber loss function or a smoothed hinge loss function:

$$g(t) = \begin{cases} 0 & \text{if} \quad t \geq 0 \\ -2t-1 & \text{if} \quad t < -1 \\ t^2 & \text{otherwise} \end{cases}, \tag{12}$$

$$g(t) = \begin{cases} 0 & \text{if} \quad t \geq 0 \\ -t-\frac{1}{2} & \text{if} \quad t < -1 \\ \frac{1}{2}t^2 & \text{otherwise} \end{cases}. \tag{13}$$

In risk function $J(\mathbf{w}, b)$, parameter λ is the regularization strength; $C \geq 0$ controls the trade-off between classification errors and ranking errors; N^+ and N^- are the number of positive and negative samples respectively. Fig. 3 illustrates the loss function used in $J(\mathbf{w}, b)$. Squared hinge loss is much more sensitive to outliers and large errors than the smoothed hinge loss and Huber loss. It is applied to patch classification to ensure the risk function is sensitive to every training label. The latter two functions are choices for the pairwise ranking errors. OPT images of colorectal polyps usually involve large intra-class variations so we expect the pairwise outliers would not dominate the risk function. There are other loss functions that meet our requirements [5]. We chose these convex and smooth functions as they can be efficiently integrated into SAG.

To apply SAG methods for minimizing $J(\mathbf{w}, b)$ iteratively, at each iteration $\mathbf{w}$ is updated with an average of the gradient of a randomly selected training pair $(\mathbf{x}_i, \mathbf{x}_j)$ and most recently computed gradients of the other training pairs. At the *(k+1)*th iteration the updating rule with a small step size α_k has the form:

$$\mathbf{w}^{k+1} = (1-\alpha_k\lambda)\mathbf{w}^k - \frac{\alpha_k}{N^+N^-} \sum_{\mathbf{x}_i \in \mathbf{X}^+} \sum_{\mathbf{x}_j \in \mathbf{X}^-} grad_{ij}^k, \tag{14}$$

where for the training pair (i^k, j^k), we set:

$$grad_{ij}^k = \begin{cases} f_i' + f_j' + Cg_{ij}' & \text{if}(i,j) = (i^k, j^k) \\ grad_{ij}^{k-1} & \text{otherwise} \end{cases}. \tag{15}$$

For the bias term b, we simply extend each feature vector with one bias component (from $\mathbf{x}$ to $[\mathbf{x}; b]$) in each iteration. This method has an exponential convergence rate and with a few implementation tricks (described in [6]) we reduce the storage cost to $\mathcal{O}(N^+N^-)$. This allows the method to scale to large datasets.

3 Experiments

Data. OPT images from 59 patients acquired using ultraviolet light and Cy3 dye were used in this study. Each image was of one colorectal polyp specimen and had $1024 \times 1024 \times 1024$ voxels with aspect ratio $1 : 1 : 1$. For each volumetric image, 3D regions were annotated by a trained pathologist with labels of dysplastic change. In 30 images, regions judged to consist entirely of low-grade dysplasia (LGD) were annotated. In the other 29 images, regions judged to consist entirely of invasive cancer were annotated. During the manual annotation process, the pathologist was asked to roughly indicate significant regions instead of exhaustively delineating all the regions.

Experimental Setup. We evaluated two aspects of the proposed model in terms of patch classification performance: (1) the ability to utilise unlabelled patches, compared with not using unlabelled patches, and using unlabelled patches naively (standard SVM); (2) the choice of loss function and affinity measurement.

In the experiments we applied 10-fold cross-validation. The dataset was randomly split into 10 folds (about 3 cancer and 3 LGD images per fold); 10 iterations of training and testing were performed such that within each iteration one fold was used as test set. The performance was averaged over the 10 iterations.

Test patches were randomly sampled from annotated regions in the test sets (about 1400 patches per fold). The partial annotation process was simulated by randomly sampling point locations within the pathologist-annotated regions. Candidate patches were randomly sampled outside the annotated regions in the training set. With reference and candidate patches, three types of models were trained:

- **T1**: training with only reference patches, using standard SVM. (*SVM-ref*)
- **T2**: training with both reference and candidate patches, using standard SVM. Labels of candidate patches can be assigned with either feature-based or location-based affinity (formula (2) or (3)). (*SVM-fea* and *SVM-loc*)
- **T3**: using our proposed model with both reference and candidate patches. We evaluated four combinations of different loss functions (formula (12) and (13)) and affinities (formula (2) and (3)). (*Prop-hub-fea*, *Prop-hub-loc*, *Prop-squ-fea*, and *Prop-squ-loc*)

We started with a training set with only 1 reference patch and 20 candidate patches for each training image (training set size: $1,140$). The models were learned using T1, T2 and T3 methods respectively and the classification performance was evaluated on the test patches. Then we added more reference patches and their associated candidate patches. At each iteration 1 reference and about 6 candidate patches per image were added. Such iterations were repeated 30 times till there were $1,590$ reference patches in the training set. At the final iteration the number of training patches was about $12,000$.

The size of each image patch was set to $21 \times 21 \times 21$ voxels. For feature extraction we used Bag of Words with Random Projection since this achieved

highest classification accuracies in [4]. The dimensionality of each feature vector was 200. Each feature was normalized to zero mean and unit variance.

In all standard SVM evaluations we used the LIBLINEAR [3] solver that solves the $L2$ regularized squared loss primal problem (with regularization parameter searched from 10^{-7} to 10^{7} and $eps = 0.01$). In our proposed method, C searched from 10^{-10} to 10^{-5}, $\lambda = \frac{1}{N^{+}N^{-}}$, $b = 0$, and the stochastic gradient step size was set to 0.004. The scaling factors were estimated from standard deviation of all distances ($\|\mathbf{z}_k - \mathbf{z}_r\|$ or $\|\mathbf{x}_k - \mathbf{x}_r\|$) between reference patches and candidate patches in the training set ($\sigma = 158.1$ in formula (2), $\delta = 7071.1$ in formula (3)).

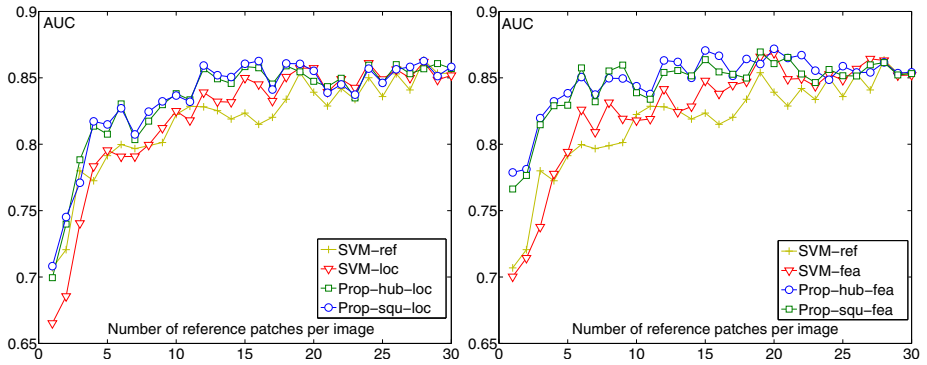


Fig. 4. AUC values depending on number of reference patches with (a) location-based affinity measurement, and (b) feature-based affinity measurement

Results. Area Under ROC Curve (AUC) obtained when classifying patches as LGD or invasive cancer was used as a performance measure. Fig. 4 shows AUC values depending on the number of reference patches per training image. We list the AUC values depending on number of reference patches per image in Table 1.

In Fig. 4, with more than 10 reference patches per image (530 reference patches, about 3,000 training patches in total) the classification performances of all the methods saturated. With both location-based and feature-based affinity the SVM-loc method showed the same or slightly higher AUCs than the SVM-ref method. This indicates that simply feeding uncertain patches to standard SVM does little to help patch classification performance. The information presented in uncertain patches was not utilized effectively by standard SVM. The proposed methods performed relatively well with small training sets indicating that they were making effective use of the unannotated patches. AUCs of all methods converged to similar values when number of reference patches reaches 30 per image.

For both affinity-based experiments, the proposed models with Huber loss and smoothed hinge loss showed almost the same AUCs. However our grid search of parameters showed that the best parameters C are quite different ($C = 10^{-4}$ for Huber loss and $C = 10^{-2}$ for smoothed hinge loss).

Table 1. Performance comparison between standard SVM and proposed model. AUC values(%) ± standard errors depending on the number of reference patches per image.

affinity	Location-based				Feature-based			
patches	1	2	5	10	1	2	5	10
SVM-ref	$\mathbf{71 \pm 2.9}$	72 ± 1.4	79 ± 2.3	82 ± 2.1	71 ± 2.9	72 ± 1.4	79 ± 2.3	82 ± 2.1
SVM	67 ± 2.2	69 ± 3.0	80 ± 2.2	83 ± 2.3	70 ± 2.8	71 ± 2.5	79 ± 2.8	82 ± 1.7
Prop-hub	70 ± 2.2	74 ± 2.0	$\mathbf{81 \pm 1.9}$	$\mathbf{84 \pm 1.6}$	$\mathbf{78 \pm 2.7}$	$\mathbf{78 \pm 2.4}$	$\mathbf{84 \pm 2.6}$	$\mathbf{85 \pm 1.9}$
Prop-squ	$\mathbf{71 \pm 2.3}$	$\mathbf{75 \pm 1.8}$	$\mathbf{81 \pm 1.8}$	$\mathbf{84 \pm 1.9}$	77 ± 3.0	$\mathbf{78 \pm 2.7}$	83 ± 2.7	84 ± 1.9

4 Conclusions

We have proposed a learning model for partially annotated images. The experiment on a dataset of 59 OPT images showed that it is able to robustly learn from patches with uncertain labels, achieving high classification accuracies while reducing the annotation effort. At the same time our model can be efficiently evaluated with only $\mathcal{O}(N^+N^-)$ in storage cost. Therefore it is suitable for high-resolution, volumetric datasets.

Acknowledgement. This work is partially supported by the Dundee Cancer Centre (DCC) Development Fund and an RSE-NSFC Joint Project (RSE Reference: 443570/NNS/INT).

References

1. Doyle, S., Monaco, J., Feldman, M., Tomaszewski, J., Madabhushi, A.: An active learning based classification strategy for the minority class problem: application to histopathology annotation. BMC Bioinformatics 12(1), 1–14 (2011)
2. Dundar, M., Badve, S., Raykar, V., Jain, R., Sertel, O., Gurcan, M.: A multiple instance learning approach toward optimal classification of pathology slides. In: ICPR, pp. 2732–2735 (2010)
3. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research 9, 1871–1874 (2008)
4. Li, W., Zhang, J., McKenna, S.J., Coats, M., Carey, F.A.: Classification of colorectal polyp regions in optical projection tomography. In: ISBI (2013)
5. Rennie, J.D., Srebro, N.: Loss functions for preference levels: Regression with discrete ordered labels. In: Proceedings of the IJCAI Multidisciplinary Workshop on Advances in Preference Handling, pp. 180–186 (2005)
6. Roux, N.L., Schmidt, M., Bach, F.: A stochastic gradient method with an exponential convergence rate for finite training sets. In: NIPS 25, pp. 2672–2680 (2012)
7. Sharpe, J., Ahlgren, U., Perry, P., Hill, B., Ross, A., Hecksher-Sørensen, J., Baldock, R., Davidson, D.: Optical projection tomography as a tool for 3D microscopy and gene expression studies. Science 296(5567), 541–545 (2002)
8. Xu, Y., Zhu, J.Y., Chang, E., Tu, Z.: Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering. In: CVPR, pp. 964–971 (2012)