LEARNING DISCRIMINATIVE LOCAL FEATURES FROM IMAGE-LEVEL LABELLED DATA FOR COLONOSCOPY IMAGE CLASSIFICATION

Siyamalan Manivannan *Emanuele Trucco*

CVIP Computer Vision and Image Processing group, School of Computing, University of Dundee, UK.

ABSTRACT

In this paper we propose a novel weakly-supervised feature learning approach, learning discriminative local features from image-level labelled data for image classification. Unlike existing feature learning approaches which assume that a set of additional data in the form of matching/non-matching pairs of local patches are given for learning the features, our approach only uses the image-level labels which are much easier to obtain. Experiments on a colonoscopy image dataset with 2100 images shows that the learned local features outperforms other hand-crafted features and gives a state-or-the-art classification accuracy of 93.5%.

Index Terms— Discriminative feature learning, Local Binary Patterns, Colonoscopy image classification.

1. INTRODUCTION

More than one million new colorectal cancer cases are diagnosed yearly worldwide. Colorectal cancer is the second leading cause of cancer death in the world and the third most common cancer in the UK [1]. Colonoscopy is the gold-standard procedure to inspect the colonic mucosa, in a relatively painless way. The *adenoma detection rate* (ADR) is a common predictor of the risk of developing colorectal cancer after undergoing a colonoscopy screening [2]. It has been argued that a reliable image processing system detecting abnormalities (including polyps, cancer, ulcers, etc.) in colonoscopy videos would be a useful screening tool to improve ADR by providing a consistent, repeatable and quantitative second opinion [3]. Here, we concentrate on normal-abnormal frame classification, a challenging task as abnormalities in colon vary in size, type, color, and shape (Figure 1). Within this task, we address the crucial aspect of *feature descriptors*.

The majority of the methods proposed for colonoscopy image classification are mainly focussed on designing appropriate features, and various hand-crafted features, such as SIFT [4], color histograms [5], Local Binary Patterns (LBP) [4], and Color Wavelet Co-variances features [6], have been explored. These features may however not optimally discriminative for classifying images from particular domains (e.g. colonoscopy) as they are hand-crafted and may not capture the particular characteristics of a specific domain.

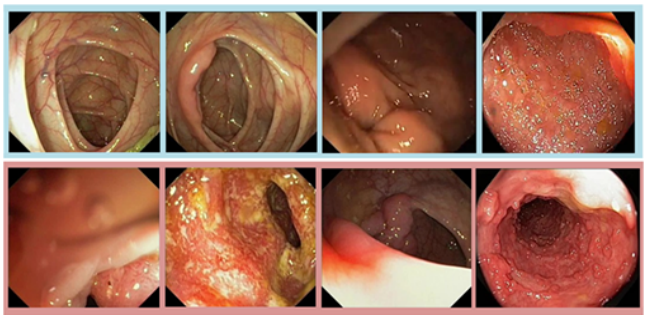

Fig. 1. Example colonoscopy images from our dataset: normal (top) and abnormal (bottom).

Recently, feature learning approaches have become popular [7, 8, 9, 10, 11] which learn domain-specific discriminative features and hence improve the performance of image classification, image retrieval, and/or interest point matching. These approaches assume that a training set consisting of matching and non-matching pairs of image patches are available to learn the feature descriptors. Such feature learning techniques have not been explored for colonoscopy image classification yet. Since obtaining matching and non-matching pairs of image patches in sufficient quantities is a difficult, time consuming task, we propose a novel weakly-supervised feature learning approach which uses instead *image-level training labels*. Requiring labels for images, not for patches makes annotations inexpensive, hence more feasible in practice.

We propose a novel, discriminative feature learning approach based on the *image-to-class (I2C) distance measure* [12]. In our approach, features are learned via a max-margin optimization formulation motivated by SVM, where we minimize the I2C distance between an image and its belonging class while maximizing the distance between that image to other classes. Note that our feature learning approach is not restricted to colonoscopy images, and can be applied to any other domain for which image-levels labels are available in a training dataset.

The contributions of this paper are:

- a novel weakly-supervised (uses image-level labelled data) feature learning approach to learn discriminative local (patch-based) feature descriptors;
- a max-margin optimization formulation based on image-to-class distances to learn the features;

- the application of the feature learning approach to colonoscopy image classification.

Comparative experiments shows that our learned features outperform widely applied feature descriptors such as LBP, SIFT, Random Projections, and color histograms.

2. METHODOLOGY

This section describes our discriminative feature learning approach. First we describe the structure of the proposed feature, and then the max-margin optimization framework created to learn the feature parameters. We call the learned feature "*Extended Multi-Resolution Local Pattern*" (xMRLP).

Let $\{I_i, y_i\}$, $i = 1, \ldots, N$ represents the training data, where I_i and $y_i \in \{1, \ldots, C\}$ are the i^{th} image and its corresponding label respectively (C represents the number of classes). I_{ij} be the intensity value of the j^{th} pixel in image I_i. To capture larger local regions (larger than 3×3 local image neighbourhood) and to make the descriptor less sensitive to noise we use a sampling pattern inspired by the spatial structure of the receptors in the human retina and has become popular in recent work on visual descriptors, e.g. BRISK [13]. Figure 2 shows a 3-resolution version of the sampling pattern, where the local neighbourhood around the j^{th} pixel of image I_i is quantized into three resolutions. At each resolution, a set of (8) sampling points are considered. At each sampling point, a Gaussian filter with standard deviation proportional to the size of the support region (circle around each sampling point in Figure 2) is applied to collect information from a region which is larger than one pixel.

Let I^s_{ij}, $s = 1, \ldots, d$, represents the intensity value at the s-th sampling point (after filtering with Gaussian as explained above) around the j^{th} pixel of image I_i (e.g. $d = 24$ in Figure 2). We define $\mathbf{x}_{ij} \in \mathbb{R}^d$ as xMRLP descriptor at pixel j in image I_i using the multi-resolution sampling pattern with d sampling points, i.e.,

$$\mathbf{x}_{ij}(\mathbf{a}, \tau) = \begin{bmatrix} I_{ij} & + & a_1 I^1_{ij} & + & \tau_1 \\ \vdots & & \vdots & & \vdots \\ I_{ij} & + & a_d I^d_{ij} & + & \tau_d \end{bmatrix} \tag{1}$$

Where, $\mathbf{a} = \{a_1, \ldots, a_d\}$ and $\tau = \{\tau_1, \ldots, \tau_d\}$ are a set of parameters which have to be learned. This local feature is motivated by the success of LBP descriptor, which is a special case of xMRLP with $a_s = -1$, and $\tau_s = 0$ for all $s = 1, \ldots, d$.

To learn the parameters $\mathbf{a}$ and τ, we propose a max-margin framework based on the image-to-class (I2C) distance. I2C distance is a non-parametric method, which was initially proposed by Bioman et al. [12] within the naive Bayes nearest-neighbour (NBNN) classifier, and then subsequently used for distance metric learning [14] and discriminative subspace learning [15]. Image classification based on the I2C distance does not require a training phase: it classifies a test image by comparing its distances to different classes, and choosing the smallest one. The relaxed version (considering more than one nearest neighbour) of I2C distance between an image I_i to a class c can be defined as (with an added normalization factor) [15]:

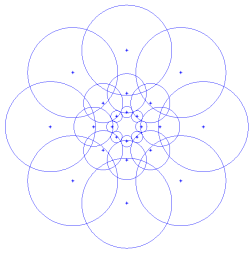

Fig. 2. A three-resolution sampling pattern.

$$D_{ic} = \frac{1}{N_i P} \sum_{j=1}^{N_i} \sum_{k=1}^{K} \|\mathbf{x}_{ij} - \mathbf{x}^{ck}_{ij}\|_2^2 \,, \tag{2}$$

where $\mathbf{x}^{ck}_{ij}$ is the k^{th} nearest neighbour of $\mathbf{x}_{ij}$ in the c^{th} class, K is the number of neighbours considered, and N_i is the number of local features in the image I_i. Note that we write $D_{ic}(\mathbf{a}, \tau) \equiv D_{ic}$ for notation compactness.

Motivated by the soft-margin loss function of SVM and by the distance metric learning framework of [14], we propose the following minimization framework to learn $\mathbf{a}$ and τ. In this approach the I2C from image I_i to its correct class c should be smaller than the distance from I_i to any other class $\bar{c}$, i.e.,

$$\begin{aligned} \arg\min_{\mathbf{a},\tau} & \sum_{c=1}^{C} \frac{1}{N_c} \left[\sum_{i \in c} D_{ic} + \lambda \xi_{ic\bar{c}} \right] \\ s.t. \quad & D_{i\bar{c}} - D_{ic} \geq 1 - \xi_{ic\bar{c}} \\ & \xi_{ic\bar{c}} \geq 0 \end{aligned} \tag{3}$$

where the non-negative slack variable ξ measures the degree of misclassification, and N_c is the number of images in class c. When we calculate D_{ic} we use the local features from all the training images except I_i. The above problem (Eqn. (3)) can be rewritten using a target functional which includes regularization terms as follows:

$$\begin{aligned} L(\mathbf{a}, \tau) = & \sum_{c=1}^{C} \frac{1}{N_c} \sum_{i \in c} \left[D_{ic} + \lambda \max\left(0, 1 - D_{i\bar{c}} + D_{ic}\right) \right] \\ & \beta \sum_{s=1}^{d} (a_s + 1)^2 \; + \; \gamma \sum_{\forall i,j} \|\mathbf{x_{ij}}\|_2^2 \end{aligned} \tag{4}$$

where the second term is a regularization term forcing the values of $\mathbf{a}$ close to -1. Since most of the image regions are smooth the last term minimizes the values of $\mathbf{x}_{ij}$.

We use an iterative gradient descent method to optimize Eqn (4). As shown in Algorithm 1, we use an alternate minimization approach, where we fix τ and optimize $\mathbf{a}$, and vice-versa. Algorithm 2 learns $\mathbf{a}$ when τ is fixed, where $\nabla_{\mathbf{a}} L$ is the

Algorithm 1 Parameter learning

Input: $\{I_i, y_i\}$, $i = 1, \ldots, N$
Output: $\mathbf{a}, \tau$

1: initialize $\mathbf{a} = [-1, \ldots, -1]^T$ and τ
2: **while** not converged **do**
3: $\quad \mathbf{a} \leftarrow$ Update $\mathbf{a}$ using Algorithm 2
4: $\quad \tau \leftarrow$ Update τ (see text)
5: **end while**

Algorithm 2 Update $\mathbf{a}$

Input: $\mathbf{a}, \tau, \{I_i, y_i\}$, $i = 1, \ldots, N$
Output: $\mathbf{a}$

1: **while** not converged **do**
2: $\quad$ compute $\mathbf{x_{ij}}$ using current $(\mathbf{a}, \tau)$ (Eqn. 1)
3: $\quad$ compute the I2C distances D_{ic} and $D_{i\bar{c}}$ (Eqn. 2)
4: $\quad$ update $\mathbf{a}$: $\mathbf{a} \leftarrow \mathbf{a} - \eta_a \nabla_{\mathbf{a}} L$
5: **end while**

gradient of Equation (4) w.r.t. $\mathbf{a}$, and η_a is the learning rate of the parameter $\mathbf{a}$. Updating τ is very similar to Algorithm 2, with line 4 featuring τ instead of $\mathbf{a}$. We initialize a_s to -1 and τ_s to small random values sampled from a Gaussian distribution (mean 0 and std 1) respectively ($s = 1, \ldots, d$).

3. EXPERIMENTS

We evaluate the proposed feature learning method with a colonoscopy image dataset containing 2100 images, half of which are normal. Abnormal images contain various kinds of abnormalities including polyps, cancers, ulcers, bleeding, etc. These images are obtained from the Internet and annotated by a clinician. All the images were rescaled preserving the aspect ratio so that the maximum dimension was 300 pixels.

For classification, we use a SVM classifier (LibSVM [16]) with the exponential chi-square kernel $K(\mathbf{H}_1, \mathbf{H}_2) = exp\left(-\frac{\gamma}{2}\sum_{i=1}^{M}\frac{(H_{1i}-H_{2i})^2}{H_{1i}+H_{2i}}\right)$, where $\mathbf{H}_1$ and $\mathbf{H}_2$ are M-dimensional representations of two images, and H_{1i} is the i-th components of H_1. The kernel parameter (γ) and the regularization parameter (C) of the soft-margin SVM are learned based on a 5-fold cross-validation on the training data.

Classification results are averaged over 10 experimental runs. In each run we randomly select 300 images from each class for training and use the rest for testing. Classification performance is measured as the mean-class-accuracy (MAC) on the test set. MAC is the average of the percentages of correctly classified images in each class.

3.1. Performance of the proposed xMRLP

To guarantee a fair comparison between descriptors, the sampling pattern in Figure 2 is first rescaled so that the sampling points in the outer most region lie on the edge of a patch of size 16×16 pixels. In the parameter learning stage we extract the xMRLP features densely (without overlap for computational efficiency) from each color channel in RGB space, and concatenate them to obtain a feature vector of size 72 (3 colors $\times$ 3 resolutions $\times$ 8 sampling points). Since the I2C calculations are computationally expensive due to the nearest neighbour search, we use only 400 images (out of 600) randomly sampled from the training set in each experimental run for parameter learning.

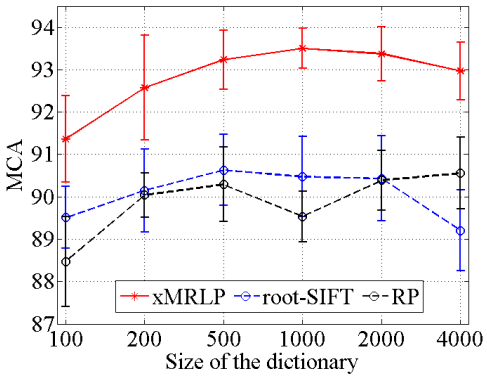

Fig. 3. Comparison of xMRLP, SIFT and RP for different dictionary sizes.

After learning the parameters $\mathbf{a}$ and τ using Algorithm 1, we extract xMRLP features densely with an overlap of 12 pixels in the horizontal and vertical directions using the parameters learned. We use a bag-of-visual-words (BOW) approach to obtain the image representations from the xMRLP descriptors: a set of $100,000$ randomly sampled xMRLP descriptors is used to learn the dictionary. Figure 3 reports the performance of the proposed feature learning approach for different dictionary sizes. The xMRLP feature gives a MCA of 93.5 ± 0.47 when the dictionary size is 1000. Increasing the dictionary size from 2000 to 4000 slightly reduces the classification performance, this may be due to the fine partition of the feature space leading to the formation of some noisy clusters, and the hard descriptor-to-cluster assignments of BoW framework.

3.2. Comparison with existing features

This section compares the proposed xMRLP descriptor with recent, popular features in image classification and computer vision: LBP [17], Local Ternary Patterns (LTP) [18], root-SIFT [19] and Random Projection (RP) [20].

LBP and LTP are texture descriptors widely applied to texture classification [17], face recognition [18], colonoscopy image classification [4], etc. Root-SIFT is an enhanced variant of SIFT, reportedly performing better than SIFT for some image retrieval tasks [19]. RP is a dimensionality reduction technique, projecting the vectorized raw patches into a compressed space [20], also adapted in medical image classification [21].

In the interest of a fair comparison, we densely extract LBP and LTP descriptors with an overlap of 12 pixels and the same sampling patterns used above (scaled to fit 16×16).

Feature	CH	CWC	CWC2	GLCM	LBP	LTP	root-SIFT	RP	xMRLP
M	225	216	240	144	531	1062	1000	1000	1000
MCA	85.1 ± 0.9	62.1 ± 1.1	62.3 ± 1.1	77.1 ± 1.2	87.2 ± 1.1	88.8 ± 1.0	90.4 ± 0.9	89.5 ± 0.5	$\mathbf{93.5 \pm 0.4}$

Table 1. Experimental results of the proposed xMRLP feature against other features (M is the size of the image representation).

In each experimental run, LTP parameters are learned in a 3-fold cross-validation on the training set. We use the uniform histogram representation of LBP as it known to capture meaningful local patterns [17]. Since we use a 3-resolution sampling pattern and extract features from 3 different color channels, the dimensionality of the image representation based on LBP is 531 (3 colors $\times$ 3 resolutions $\times$ 59, where 59 is the dimensionality of the uniform LBP histogram obtained from a single resolution and monochromatic channel). The LTP dimensionality is 1062 (double that of LBP). It can be seen from Table 1 that the learned xMRLP descriptor outperforms the LBP and LTP descriptors.

The root-SIFT features extracted (patch size 16×16, overlap 12 pixels) from 3 color channels are concatenated to get a descriptor of size 3×128 . RP descriptors are formed in a similar manner; The concatenated vectorized patches ($3 \times 16 \times 16$) are projected by a RP matrix to obtain a feature representation of size 200 for each patch. We use BOW for feature encoding. Figure 3 reports the performance of root-SIFT and RP descriptors for different dictionary sizes. Root-SIFT performs similar to RP, and the performance of both are worst than that of the proposed xMRLP descriptor.

Table 1 also compares xMRLP descriptors with the features adapted in the recent colonoscopy (*white-light*) image classification literature, such as color histograms (CH) [5], color wavelet covariance (CWC)[6], CWC with higher-order statistics (CWC2)[22] and Gray Level Co-occurrence Matrices (GLCM)[23]. Our results show that all these descriptors are outperformed by xMRLP.

4. CONCLUSIONS

We presented a novel discriminative feature learning approach using image-level labels to learn discriminative local features. We introduced a novel descriptor, xMRLP, and showed in comparative experiments that our learned feature outperform recent popular descriptors from the computer vision as well as the colonoscopy image classification literature. Future work will explore the applications of the proposed feature learning method to classification of further domains of medical images.

Acknowledgement: Work is funded by 2011-2016 EU FP7 ERC "CODIR: colonic disease investigation by robotic hydrocolonoscopy", Universities of Dundee (PI Prof Sir A Cuschieri) and Leeds (PI Prof A Neville). We thank Dr. Adrian Hood for annotations, and Dr. Ruixuan Wang for useful discussions.

References

[1] "Cancer research uk," `info.cancerresearchuk.org/cancerstats`.

[2] Wallace MB, "Improving colorectal adenoma detection: technology or technique?," *Gastroenterology*, 2007.

[3] MF Madhoun and WM Tierney, "The impact of video recording colonoscopy on adenoma detection rates," *Gastro. Endoscopy*, 2012.

[4] S. Manivannan, R. Wang, E. Trucco, and A. Hood, "Automatic normal-abnormal video frame classification for colonoscopy," in *IEEE ISBI*, 2013.

[5] P.C. Khun et al., "Feature selection and classification for wireless capsule endoscopic frames," in *Int. Conf. on Biomed. and Pharmaceutical Eng.*, 2009.

[6] S.A Karkanis, D.K Iakovvidis, D.E Maroulis, D.A Karras, and M Tzivras, "Computer aided tumor detection in endoscopic video using color wavelet features," *IEEE trans. on IT in biomedicine*, 2003.

[7] M. Brown, Gang Hua, and S. Winder, "Discriminative learning of local image descriptors," *IEEE Trans. on PAMI*, 2011.

[8] K. Simonyan, A. Vedaldi, and A. Zisserman, "Learning local feature descriptors using convex optimisation," *IEEE Trans. on PAMI*, 2014.

[9] Simon A. J. Winder and Matthew Brown, "Learning local image descriptors," in *IEEE CVPR*, 2007.

[10] Philbin J., Isard M., Sivic J., and Zisserman A., "Descriptor learning for efficient retrieval," in *ECCV*, 2010.

[11] Winder S, Hua G, and Brown M, "Picking the best daisy," in *IEEE CVPR*, 2009.

[12] O.Boiman, E.Shechtman, and M.Irani, "In defense of nearest-neighbor based image classification," in *IEEE CVPR*, 2008.

[13] S. Leutenegger, M. Chli, and R.Y. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *IEEE ICCV*, 2011.

[14] Z.Wang, Y.Hu, and L.T. Chia, "Image-to-class distance metric learning for image classification," in *ECCV*, 2010.

[15] X. Zhen, L.Shao, and F. Zheng, "Discriminative embedding via image-to-class distances," in *BMVC*, 2014.

[16] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM: A library for support vector machines," *ACM Trans. on Intelligent Sys. and Tech.*, 2011.

[17] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. on PAMI*, 2002.

[18] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE TIP*, 2010.

[19] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *IEEE CVPR*, 2012.

[20] Ella Bingham and Heikki Mannila, "Random projection in dimensionality reduction: Applications to image and text data," in *ACM KDDM*, 2001.

[21] S Manivannan, W Li, S Akbar, R Wang, J Zhang, and S J. McKenna, "Hep-2 cell classification using multi-resolution local patterns and ensemble svms," in *workshop on PR Techniques for IIF images, ICPR*, 2014.

[22] C. S. Lima, D. Barbosa, A. Ramos, A. Tavares, L. Montero, and L. Carvalho, "Classification of endoscopic capsule images by using color wavelet features, higher order statistics and radial basis functions," in *EMBS*, 2008.

[23] Sandy Engelhardt, Stefan Ameling, Dietrich Paulus, and Stephan Wirth, "Features for classification of polyps in colonoscopy," 2010, CEUR Workshop Proceedings.