Automatic Polyp Frame Screening using Patch based Combined Feature and Dictionary Learning

Younghak Shin and Ilangko Balasingham, *Senior Member, IEEE*

***Abstract*—Polyps in the colon can potentially become malignant cancer tissues where early detection and removal lead to high survival rate. Certain types of polyps can be difficult to detect even for highly trained physicians. Inspired by aforementioned problem our study aims to improve the human detection performance by developing an automatic polyp screening framework as a decision support tool. We use a small image patch based combined feature method. Features include shape and color information and are extracted using histogram of oriented gradient and hue histogram methods. Dictionary learning based training is used to learn features and final feature vector is formed using sparse coding. For classification, we use patch image classification based on linear support vector machine and whole image thresholding. The proposed framework is evaluated using three public polyp databases. Our experimental results show that the proposed scheme successfully classified polyps and normal images with over 95% of classification accuracy, sensitivity, specificity and precision. In addition, we compare performance of the proposed scheme with conventional feature based methods and the convolutional neural network (CNN) based deep learning approach which is the state of the art technique in many image classification applications.**

***Index Terms*— Colonoscopy, computer-aided detection, shape and color feature, dictionary learning, polyp classification, sparse coding.**

I. Introduction

Colorectal cancer (CRC) is the third leading cancer to cause deaths in United States in both men and women. In 2016 a total of 49,190 deaths were due to CRC [1]. CRC arises from adenomatous polyps. These colonic polyps are growths of glandular tissue at colonic mucosa. Even though most polyps are initially benign, they can be malignant over time. Therefore, detection of polyps in their early stage is very important to prevent the CRC [2][7].

Colonoscopy (or endoscopy) is considered as gold standard for screening polyps. Studies show that colonoscopy successfully contributed to a 30% decline in the incidence of CRC [3]. However, conventional polyp detection using colonoscopy is fully operator dependent procedure where polyp miss-detection rate is known as about 25% [4]. The miss-detected polyps can lead to late diagnosis of CRC having 10% survival rate [5]. Miss-detection seems due to lack of notable characteristics for human vision and insufficient attentiveness of clinician during long colon examination [13].

Recently, wireless capsule endoscopy (WCE) has been proposed to avoid discomfort associated with colonoscopy examination [2][6]. WCE, an electronic pill, consists of image sensor, LED lights, wireless transmitter, and battery. When swallowed by a patient it travels using the peristaltic movement in the gastrointestinal tract. The transmitted images are received by an antenna array placed around the waist of the patient. WCE has advantages that patients can avoid cross infection and suffer no pain. However, the captured images for one patient are known as over 50,000 frames and it needs long time examination by the experienced physician to find abnormalities [2]. Therefore, automatic polyp detection methods are helpful to improve the human detection performance. Due to there is no public WCE polyp dataset available, this study therefore focuses only on analysis of normal colonoscopy datasets.

Over last two decades, many strategies for computer-aided detection (CAD) of polyp are proposed to increase polyp detection rate and to reduce time costs of clinicians [7]-[15]. In the initial stages, color and texture based features such as color wavelet and local binary pattern (LBP) are used for detection of polyp region [7][8][39]. In [9], elliptical shape fitting based polyp detection method is suggested. However, these methods introduce similar feature pattern between polyp and non-polyp part that causes performance decrease.

More recently Bernal *et al.* proposed a polyp localization method by modeling polyp appearance using the valley information of polyp boundaries [10][11]. In [45], image part based edge cross-section profiles was used for precise polyp detection. An imbalanced learning and discriminative feature learning scheme was proposed for a balanced training between polyp and non-polyp images [12]. In [13], they used shape and context information to improve discriminative power of polyp with other polyp-like structures. However, due to a large intra-class variation of polyp appearances, hand-craft feature type based polyp localization is a difficult task.

With the recent success of deep learning in many image recognition applications, convolutional neural network (CNN)

This work was supported by the European Research Consortium for Informatics and Mathematics (ERCIM) 'Alain Bensoussan' Fellowship Programme and Research Council of Norway through the MELODY project under the contract number 225885/O70.

Y. Shin is with the Department Electronics and Telecommunications at Norwegian University of Science and Technology (NTNU), Trondheim, Norway (e-mail: shinyh0919@gmail.com).

I. Balasingham is with the Intervention Centre, Oslo University Hospital, Oslo NO-0027, Norway (phone: +47-230-70101; fax: +47-230-70110; e-mail:ilangkob@medisin.uio.no). He is also with the Institute of Clinical Medicine, University of Oslo, and the Norwegian University of Science and Technology (NTNU).

based deep learning feature has been proposed for polyp detection [14][15]. Furthermore, [44] reports that several teams used CNN approaches for the polyp detection challenge in 2015 international conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). In [48], pre-trained CNN with a proper fine-tuning approach is used for polyp detection. These methods focus on the automatic polyp localization– find the exact position of polyps within an image. However, precise polyp localization seems very challenging due to a variety of types of polyp appearances, i.e., in terms of size, shape, texture and color. In addition, changing of camera viewpoint, lighting condition and reflection during the colonoscopy are major obstacles for polyp localization. As we will discuss in Section III-F, polyp miss-detection rate of recent automatic polyp localization studies is still high, and it will be a limiting factor to be considered in a clinical colonoscopy diagnostic system.

Alternatively, in this study, we focus on computer-aided polyp screening framework. Thus, we aim to classify image frames which include polyp parts from normal (non-polyp) image frames. We expect that higher classification performance of the computer-aided polyp screening will have advantages to the clinician to improve performance in colonoscopy diagnosis. Figure 1 shows the concept of the computer-aided polyp screening within the normal colonoscopy diagnosis procedure by a skilled clinician. This Computer-aided polyp screening can be helpful to find missed polyp image frames by the clinician. This will help to reduce the polyp miss-detection rates. Using the automatic polyp screening framework, the skilled clinician can focus more on the automatically classified polyp image frame instead of the whole video and find the exact location of polyps without too much effort. This will facilitate faster diagnosis leading to remove all polyps.

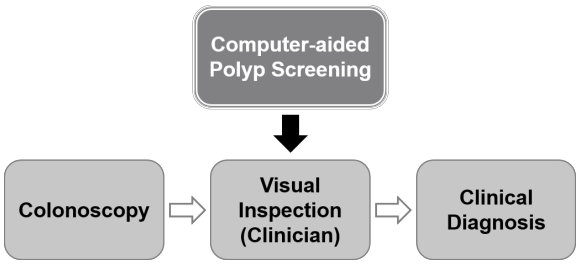

Figure 1. Computer-aided polyp screening in clinical colonoscopy procedure

Sparse representation (or sparse coding) has received a lot of attention in various signal processing fields [16]-[19]. In the sparse representation, signal is represented by linear combination of a few atoms in a dictionary using L1 minimization algorithms. This compact representation has successfully applied and shown good performance in many image processing fields such as image denoising [16], super-resolution [17], compressive sensing [18], and classification [19]. For some image signal, predefined dictionary such as wavelet [20] can be used. However, learned dictionary from input images by using the *dictionary learning* scheme has shown to provide improved results in many applications [21][22][23]. The dictionary learning scheme has been also applied into polyp and lung nodule detection using CT (computed tomography) image dataset [24]. Moreover, in [6], dictionary learning scheme is applied to wireless capsule endoscopy (WCE) image classification. However, they demonstrated only organ classification instead of polyp classification.

In this study, we propose a combined feature based dictionary learning scheme for automatic polyp screening method. To capture polyp characteristics precisely, we use a sliding window based image patch method and combined feature scheme using shape and color features. HOG (histogram of oriented gradient) and hue histogram method are used for each shape and color feature extraction. In addition, dictionary learning scheme is used for fitting the combined feature of training images and final feature is formed by sparse coding using the learned dictionary. We also suggest a two-step classification scheme, i.e., patch based classification using a linear support vector machine (SVM) and whole image classification by a simple threshold based method. For reliable evaluation, we assign different datasets for training and testing using three public colonoscopy datasets. Thus, training and test datasets are obtained separately by different patient, date, and recording system. We compare classification performance of the proposed scheme with the CNN based deep learning approach using same image patch based polyp screening framework. Furthermore, we examine classification performance of the polyp localization studies where same colonoscopy datasets were used with this study.

The rest of the paper is organized as follows. In Section II, proposed whole framework and each methodological step are introduced. In Section III, image datasets, experimental setup and results are explained. Furthermore, we provide some important discussions. Finally, we conclude this study in Section IV.

II. Methodology

In this study, we propose a combined feature based dictionary learning scheme for polyp image classification. Image patches are obtained via sliding window and used for feature extraction. The whole classification procedure is shown in Figure 2. For training, we make a combined feature including shape and color information of endoscopy image patch. Then, we use dictionary learning and sparse coding, where a SVM classifier is trained for patch classification. For testing, same feature extraction and sparse coding step are performed by using the learned dictionary from the training part. Then, SVM classification step is executed for each test image patch. Finally, threshold based whole image classification is performed to identify polyp existence inside the whole image frame. Each part of the whole procedure is explained in the following subsections.

A. Image Preprocessing and Patch Extraction

For each endoscopy image which includes polyp or non-polyp (normal), we first remove the black background area in the four corners of the image frame. This area is not related with

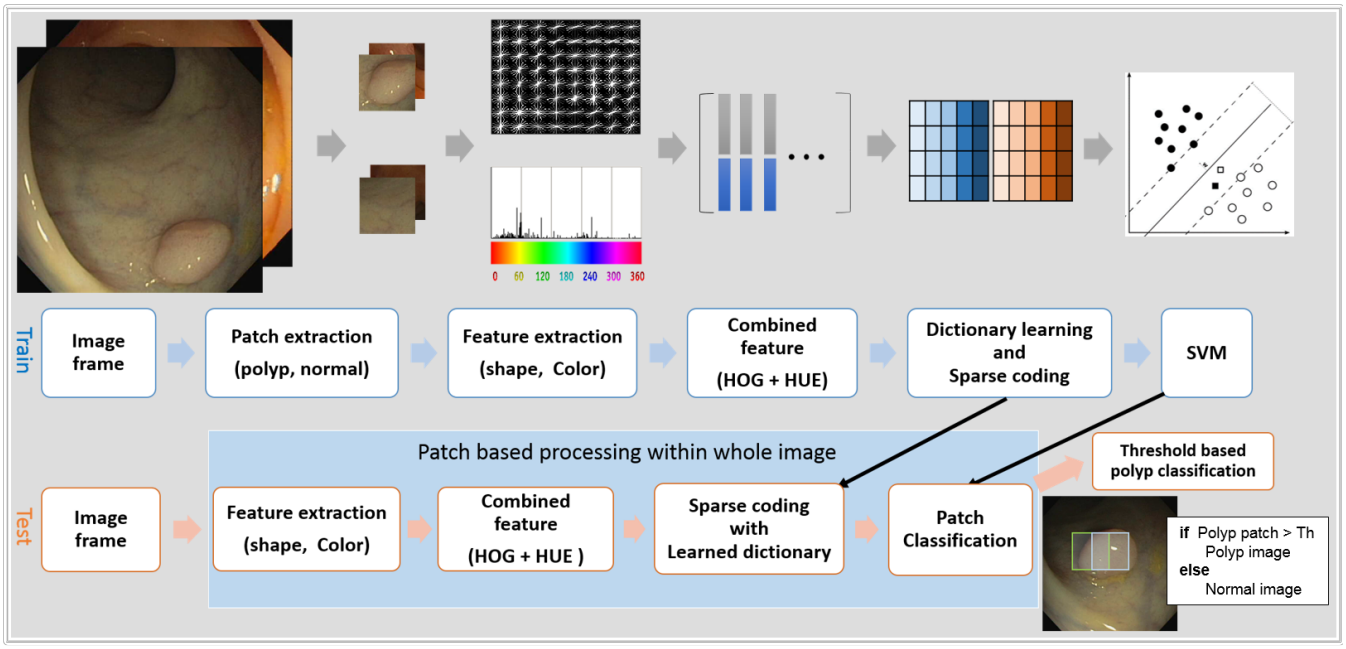

Figure 2. Proposed polyp screening framework

captured endoscopic image from inside a human organ and will not affect the next steps. Since images obtained from different databases (see Section III-*A*) had different sizes, we resized the spatial resolution to 384×384 pixels. For patch based image processing, we extract image patches from each image frame using a sliding window technique. Due to the different size and position of polyps in each image frame, sliding window is an efficient method for polyp extraction. The size of each patch is set to 128×128 where the sliding factor is 64. This means for each endoscopy image frame 25 image patches are extracted.

From the training image set, polyp image patches which include polyp parts are collected for further processing. Polyp parts are determined by the ground truth of the polyp image dataset (see Section III-*A* for detailed information). For normal image patches, except extracted polyp parts, non-polyp parts are only collected for normal training set. From the extracted polyp and normal image patches, some poor quality image patches due to the very low image resolution or strong light reflection were excluded from the training set. Some image patches including very small parts of the polyp were also excluded. From the test image set, same patch extraction step is performed for further feature extraction and classification tasks.

B. Feature extraction

In this study, *HOG (histogram of oriented gradient)* and *hue histogram* method is used as each shape and color feature extraction respectively and both are combined for further processing. We use HOG feature to capture polyp characteristics effectively. HOG feature (or descriptor) is well known shape-feature type in various object recognition fields due to its robustness against noise and illumination change of the object [25][26]. Recently, HOG feature is also applied to the polyp detection problems [12][27].

HOG feature was initially proposed by Dalal and Triggs for a specific human detection application [25]. The HOG feature is computed based on the image gradient $\nabla f(x, y)$ of each pixel point (x, y) by using simple derivative mask [-1, 0, 1] and is intended to capture the typical edge structure of the local shape. Using magnitude $\|\nabla f(x, y)\|$ and orientation $\theta(x, y)$ of the image gradient, histogram for orientation is calculated in local image cell. Here, the orientation is quantized to n bin size within ranges of $0^\circ - 180^\circ$. Finally, overlapped square block which consists of 2×2 cells is used for normalization purpose. This means for nine bin (n = 9) histogram, 36-dimensional HOG feature is extracted for each cell. For obtaining HOG feature, the bin size n and cell size s are pre-determined parameters.

Second, we consider *HSV (Hue, Saturation and Value)* color space for color feature extraction of colonoscopy image. Hue represents the base color, Saturation specifies the purity of the color, and Value represents the brightness of the color [28]. HSV color space is widely used in various image processing applications such as image segmentation [29], eye localization [30], face detection [31] and also endoscopy image processing [32]-[34]. HSV model is known as less sensitive to illumination variations compared to the RGB model [33].

In HSV model, hue, i.e., base color information, can be separated from the colorfulness and brightness [34]. Similarly, [35] shows that hue component is invariant to viewing orientation, illumination direction, intensity and highlight. As we mentioned in Section I, illumination and view point change are one of the main obstacles for polyp classification during the colonoscopy recording. Therefore, in this study, we focus on the Hue component as a robust color feature extraction. The Hue component describes the main color information in the form of an angle between [0, 360] degrees. 0, 120 and 240 degree represents red, green and blue color respectively. For each image patch, we obtained HSV color components from the RGB model by simple transformation [30]:

$$
\begin{aligned}
V &= \max(R,G,B) \\
S &= \begin{cases} 0, & \text{if } V = 0 \\ \dfrac{(V - \min(R,G,B)) \times 255}{V}, & \text{o.w.} \end{cases} \\
H &= \begin{cases} \dfrac{(G - B \times 60)}{S}, & \text{if } V = R \\ 180 + \dfrac{(B - R \times 60)}{S}, & \text{if } V = G \\ 240 + \dfrac{(R - G \times 60)}{S}, & \text{if } V = B \end{cases} \\
H &= H + 360 \quad \text{if } H < 0,
\end{aligned} \tag{1}
$$

where, H, S and V represents each Hue, Saturation and Value component respectively. The value of H is ranged from 0 to 360. After transformation, we compute a hue histogram with 16 bins for each image pixel to form a final color feature.

C. Combined Feature

Here, we examine combined feature to capture polyp characteristics effectively. We use shape and color feature by combining above mentioned HOG and hue histogram features. As shown in Figure 2, for each polyp and normal image patch, HOG and hue histogram features are separately extracted. Then, both features are concatenating into a single vector. This means the dimension of one feature vector is simply the dimension of the HOG feature plus dimension of the hue histogram feature. The combined features obtained by the training image patches are used as input for further dictionary learning process.

Figure 3 shows the example of polyp and normal part in the colonoscopy polyp images. Polyp parts represented by red line exhibit unique characteristics in terms of shape, i.e., polyp has an elliptical shape. However, some parts represented in blue lines in Figure 3, show the similar shape of the polyp part. Especially, blue line in right figure of Figure 3 is the colon lumen which has the similar elliptical shape with polyp. It can be a major cause for polyp miss-classification if we only use shape based feature. Therefore, we expect that the combined feature of shape and color information is robust for polyp classification task.

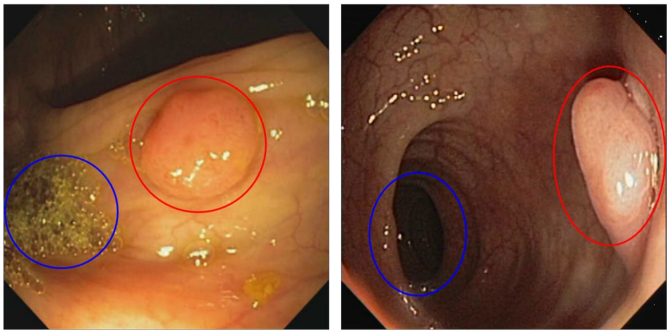

Figure 3. Similar shape between polyp and non-polyp parts in colonoscopy images

D. Dictionary Learning for Sparse Coding

Sparse coding exhibits efficiency capturing learned features from a dictionary which is learned by input training data. Concretely, using the dictionary learning process, atoms (columns) of the dictionary are adapted to input combined feature of the patch image. Then, sparse coding step is performed with the learned dictionary to form a final feature vector. This section introduces the dictionary learning and sparse coding steps.

Let $\mathbf{X}$ is a m-dimensional training feature (combined features) set: $\mathbf{X} \in \mathbb{R}^{m \times n} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$. We consider following optimization problem for dictionary learning procedures:

$$
\min_{\mathbf{D},\mathbf{h}} \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{2} \left\| \mathbf{x}_i - \mathbf{D}\mathbf{h}_i \right\|_2^2 + \lambda \left\| \mathbf{h}_i \right\|_1 \right), \tag{2}
$$

where $\mathbf{D}$ is the dictionary in $\mathbb{R}^{m \times k}$, $\mathbf{h}$ is the k-dimensional coefficient vector, and λ is a regularization parameter. Also, there is a common constraint that each column of $\mathbf{D}$ to have unit norm for prevention of large $\mathbf{D}$ in (1). Note that equation (1) is not jointly convex with $\mathbf{D}$ and $\mathbf{h}$ simultaneously, but convex with respect to each of them when the other one is fixed. Therefore, most algorithms for dictionary learning process [21]-[23] follow the alternate way, i.e., optimize $\mathbf{h}$ for fixed $\mathbf{D}$ and update $\mathbf{D}$ for obtained $\mathbf{h}$.

First, for the sparse optimization part, i.e., optimization of $\mathbf{h}$ for fixed $\mathbf{D}$, we use a *homotopy* algorithm to solve the following unconstrained optimization problem

$$
\min_{\mathbf{h}} \frac{1}{2} \left\| \mathbf{x}_i - \mathbf{D}\mathbf{h}_i \right\|_2^2 + \lambda \left\| \mathbf{h}_i \right\|_1 . \tag{3}
$$

For initial fixed $\mathbf{D}$, random initial values or k columns from input training features can be used [23]. Homotopy method is one of the greedy algorithms in sparse signal recovery. This approach is well known for elegant solution paths and fastness with a higher accuracy in many applications [23][36][37]. Therefore, in this study, we adapt the homotopy method to obtain sparse vector $\mathbf{h}$. The homotopy algorithm iteratively traces the final solution starting from the initial point by successively adjusting the homotopy parameter λ. Thus, when $\lambda \to \infty$, the solution set $\mathbf{h}_\lambda^* = 0$ and $\mathbf{h}_\lambda^*$ converging to the solution of (2) when $\lambda \to 0$. For mathematical details and implementation of homotopy algorithm, please see the reference [36][38].

Second, after obtaining $\mathbf{h}$ by the homotopy method in (2), (1) can be rewritten as follows with the unit norm constraint of $\mathbf{D}$

$$
\min_{\mathbf{D}} \frac{1}{2} \left\| \mathbf{x}_i - \mathbf{D}\mathbf{h}_i \right\|_2^2 \quad s.t. \left\| d_j \right\|_2 = 1,\ j = 1,2,\ldots,k \tag{4}
$$

where, d_j is the each column of the dictionary $\mathbf{D}$. Equation (3) can be solved by conventional gradient descent method. However we use the efficient block coordinate decent algorithm proposed in [23]. This method is known as parameter-free and does not require any learning rate tuning. For each $\mathbf{x}_i$, each column, i.e., d_j (block), of the dictionary $\mathbf{D}$ is sequentially updated by the block coordinate decent algorithm [23]:

$$
d_j \leftarrow \frac{1}{\mathbf{A}_{j,j}} \left(\mathbf{B}_j - \mathbf{D}\mathbf{A}_j \right) + d_j , \tag{5}
$$

where $\mathbf{A}_i \leftarrow \mathbf{A}_{i-1} + \mathbf{h}_i \mathbf{h}_i^T$, $\mathbf{B}_i \leftarrow \mathbf{B}_{i-1} + \mathbf{x}_i \mathbf{h}_i^T$.

Finally learned dictionary **D** is obtained by alternating above mentioned sparse optimization and dictionary update steps for each $\mathbf{x}_i$ until **D** is converged. In this study, we set the number of iteration *t* as 100 to find the converged dictionary.

We summarize the dictionary learning algorithm as follows:

Input: training feature $\mathbf{X} \in \mathbb{R}^{m \times n}$, initial dictionary $\mathbf{D}_0 \in \mathbb{R}^{m \times k}$, $\mathbf{A}_0 = 0, \mathbf{B}_0 = 0$.

1. **while** **D** hasn't converged (1, 2,…, *t*)
2. **for** $i = 1$ to n
3. Compute the $\mathbf{h}_i$ in (2) using homotopy.
4. $\mathbf{A}_i \leftarrow \mathbf{A}_{i-1} + \mathbf{h}_i \mathbf{h}_i^T$, $\mathbf{B}_i \leftarrow \mathbf{B}_{i-1} + \mathbf{x}_i \mathbf{h}_i^T$.
5. Update **D** using block coordinate decent algorithm in (4).
6. **end for**
7. **end while**

Output: learned dictionary **D**

Using a learned dictionary **D**, new $\mathbf{h}_i$ can be obtained by sparse coding of each $\mathbf{x}_i$ and the new $\mathbf{h}_i$ is a form of a final feature vector for further training process of the SVM classifier. We expect that the learned dictionary **D** which is built from the combined feature of training image patches can be considered as important basis of the input training features. For any new test feature $\mathbf{x}_{new}$, $\mathbf{h}_{new}$ is obtained by the sparse coding step with the same learned **D** and $\mathbf{h}_{new}$ is an input test feature vector.

E. Polyp Classification using Whole Image Frame

Regardless of whether it is polyp or normal test image frame, we apply the same preprocessing and patch extraction steps are explained in Section II-*A*. This means 25 image patches are extracted for each test image. For each image patch, combined feature $\mathbf{x}_{new}$ is obtained and transformed into a final feature vector $\mathbf{h}_{new}$ using the sparse coding step with the learned dictionary. The $\mathbf{h}_{new}$ for each image patch is then classified by the trained SVM. SVM is a well-known classification method in various applications of pattern recognition field including polyp detection [2][39]. SVM is recognized for its excellent generalization performance, i.e., small error rate for test data. In this study, we use a linear SVM trained with a MATLAB SVMtrain algorithm [40] and set the default regularization parameter as $C = 1$.

After image patch-classification, we introduce a threshold value (represented as Th in Figure 2) for the purpose of whole image frame-classification task. From those 25 image patches, polyp parts may appear at least in one patch within polyp image frame. On the contrary, for the normal image frame, normal parts should be in all patches ideally. However, due to the use of sliding window (polyp can be extracted in a few patches), size of the polyp and noise (any undistinguishable part between polyp and normal parts) the threshold value for the whole image classification need to be set.

Figure 4 shows an example of polyp image frame (left figure) and the extracted 25 image patches (right figure) by the sliding window. We observe that the polyp parts are shown in a few patches of the central part in the right figure. We will discuss more about the relationship between varying the threshold value and classification performance in later Section III-*E*.

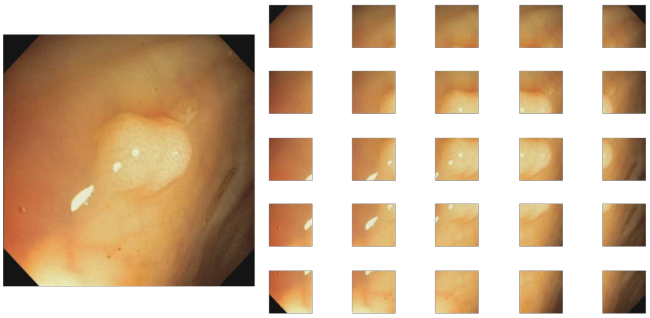

Figure 4. Example polyp image and extracted image patches using sliding window

III. Experimental Results and Discussion

A. Experiment

To evaluate the proposed polyp screening framework, we use three public datasets [42], CVC-Clinic [11], ETIS-Larib [41] and Asu-Mayo database [13]. We understand that there are similar image frames within the same colonoscopy dataset. If we use one dataset and divide that into training and testing sets, then exaggerated classification performance can be obtained. Therefore, for more reliable evaluation, we assign above mentioned different datasets into training and testing set separately. Thus, CVC-Clinic dataset is used as the training set. ETIS-Larib and Asu-Mayo datasets are used for each polyp and normal testing sets.

CVC-Clinic dataset consists of 612 polyp endoscopy images with 576×768 pixel resolution. All the images were extracted from 31 different videos and contained at least one polyp. The position of polyp part in each image is provided by expert video endoscopists from the corresponding associated clinical institution. ETIS-Larib dataset comprises 196 polyp images which are generated from 34 videos. The size of each image is 1225×966. This dataset contains 44 different polyps with various sizes and appearances. Asu-Mayo dataset comprises ten different recording times of colonoscopy videos and consists of normal (without polyp) frames. From ten videos, to avoid similar image frame, we extract 170 images using down sampling. The size of each image is 712×480.

Training set consists of 561 polyp patches and 964 normal patches extracted from CVC-Clinic dataset by using the patch extraction step explained in Section II-*A*. Testing set consists of 196 polyp images from ETIS-Larib dataset and 170 normal images from Asu-Mayo dataset.

To evaluate classification performance, we use some widely used statistical parameters such as sensitivity (recall), specificity, precision, and classification accuracy. The definition of each parameter is as follows:

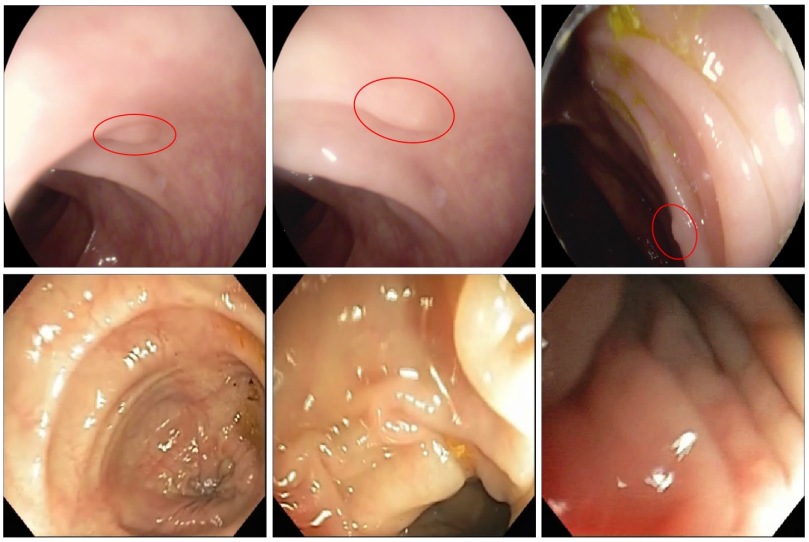

Figure 5. Example test images for miss-classification cases from the results of proposed method

Sensitivity (Recall): number of correctly classified polyp images/ total number of actual polyp images.

Specificity: number of correctly classified normal (non-polyp) images/ total number of actual normal images.

Precision: number of correctly classified polyp images / total number of classified polyp images (test results are polyp).

Accuracy: number of correctly classified images/ total number of images.

B. Experimental results

We first evaluate the detailed classification results of the proposed scheme by using the confusion matrix. Confusion matrix summarizes the classification performance in terms of actual and predicted results of each polyp and normal test image sets. Note that using the confusion matrix we can also obtain above mentioned statistical parameters.

Table 1 lists the confusion matrix for two-class (polyp/ normal) classification problem using the proposed scheme. The results show that 188 actual polyp images among 196 ETIS-Larib test images are correctly classified as polyp images by the proposed method. The number of miss-classification case, i.e., classified as normal images, is 8. In the case of normal test images from 170 Asu-Mayo dataset, 163 images are correctly classified as normal images. Only 7 images are classified as polyp images. Overall classification accuracy is 95.9 % which is obtained by (188+163)/366. Other statistical classification performances, i.e., sensitivity, specificity and precision are given in the last row in Table 2.

For the results of Table 1, cell size (s) of 64 and bin size (n) of 12 are used for HOG feature extraction, and 70 dictionary size is used for dictionary learning process. Furthermore, the threshold value (Th in Figure 2) of 3 is set for whole image classification. We will discuss more about the effect of these HOG parameters, dictionary size, and the threshold value in Section II-*C*, *D* and *E* respectively.

TABLE 1. CONFUSION MATRIX FOR POLYP CLASSIFICATION RESULTS USING PROPOSED SCHEME

		Predicted		Total
		Polyp	Normal	
Actual	Polyp	188	7	195
	Normal	8	163	171
Total		196 (ETIS)	170 (MAYO)	366
Overall accuracy: 95.9%				

In Figure 5, we show some examples of miss-classification cases from the results in Table 1. The top figure shows the three examples among eight miss-classification cases in Table 1. In this case, the actual label of these images is polyp and exact location of each polyp is represented by the red circle in Figure 5. However, as we can see, it is very difficult to find exact polyp parts in terms of shape and color. Therefore, they are miss-classified as normal images by the proposed scheme. On the other hand, in bottom figure, some actual-normal images (but classified as polyp) are shown. These images are example among seven miss-classification cases represented in Table 1. For these images, though the actual label is normal, we can see some parts show more similar shape and color with polyp than normal parts, and they are miss-classified by the proposed scheme. However, these images exhibit a reasonable doubt about the polyp existence. In addition, it will be helpful to clinician to decrease polyp miss-detection rate by carefully examining in the colonoscopy diagnosis.

For evaluation of the proposed combined feature with dictionary learning scheme, we compare classification performance of proposed framework with other different feature types in Table 2. Each color feature by the hue histogram, shape feature by the HOG, and combined color and shape feature by both methods are compared with the proposed whole framework classification.

TABLE 2. COMPARISON OF CLASSIFICATION PERFORMANCE FOR DIFFERENT METHODS

Methods	Sensitivity	Specificity	Accuracy	Precision
Color feature (Hue histogram)	0	0.9765	0.4536	No
Shape feature (HOG)	0.7041	0.8176	0.7568	0.8166
Combined feature	0.8673	0.8118	0.8415	0.8416
Dictionary learning (Homotopy) with Combined feature	0.9592	**0.9588**	**0.9590**	**0.9641**
Dictionary learning (BP) with Combined feature	0.8469	0.9059	0.8743	0.9121
Dictionary learning (FISTA) with Combined feature	**0.9847**	0.9235	0.9563	0.9369

TABLE 3. COMPARISON OF CLASSIFICATION PERFORMANCE FOR DIFFERENT PARAMETERS OF HOG FEATURE

Cell size (s), Bin size (n)	Sensitivity	Specificity	Accuracy	Precision
8, 9	0.6173	0.9059	0.7513	0.8832
16, 9	0.8214	0.8765	0.8470	0.8846
32, 9	0.8571	0.9059	0.8798	0.9130
64, 9	**0.9847**	0.9235	0.9563	0.9369
128, 9	0.9388	0.7471	0.8497	0.8106
64, 3	0.9031	0.7529	0.8333	0.8082
64, 6	0.9541	0.9118	0.9344	0.9257
64, 12	0.9592	**0.9588**	**0.9590**	**0.9641**
64, 15	0.8776	0.9000	0.8880	0.9101

For a fair comparison, we set the same threshold value 3 and use a linear SVM for patch classification in all methods. In addition, we use the same HOG parameters such as cell size (s) of 64 and bin size (n) of 12, for all HOG based methods, and a dictionary size of 70 (same as in Table 1) for dictionary learning process.

The results show combined features have improved classification performance than independent feature based method for color and shape. This is due to the color and shape features are in mutually complementary, where a combined feature seems beneficial. Moreover, the proposed dictionary learning scheme with combined feature have shown superior classification performance. All four parameters of the classification performance, i.e., sensitivity, specificity, accuracy, and precision are over 95%. This shows that the dictionary learning and sparse coding steps efficiently train and capture the input combined features and provide important foundation for classification.

In our dictionary learning process, the L1 minimization algorithm has the role to solve the sparse optimization of training data so that important training features are adapted in the updated dictionary. As mentioned in Section II-D, we adopt the homotopy algorithm for sparse optimization. However, different L1 minimization algorithms can be used for the sparse optimization purpose. Therefore, we compare classification performance of the dictionary learning for different L1 minimization algorithms in the last three rows of Table 2. We consider one widely used general purpose basis pursuit (BP) algorithm based on a standard linear programming method [18][36] and one advanced L1 minimization method which is called fast iterative shrinkage-thresholding algorithm (FISTA) [36][49]. The results indicate that regardless of L1 minimization algorithms, the dictionary learning framework shows improved classification accuracy than the without dictionary learning method. The efficient fast L1 minimization algorithms, i.e., homotopy and FISTA, outperform the BP algorithm within the dictionary learning framework and this result is consistent with other sparse representation applications in the literature [36][49].

We also evaluate the running time of the proposed method. We use the MATLAB commands tic and toc for measuring the start and stop time of the algorithm. For dictionary learning of training dataset, it takes 664.4 sec for 100 iterations. We also compute a classification time for each test image frame and averaged over all test images. The mean classification processing time is about 0.095 sec per frame (0.0038 sec per

image patch). Note that the running time of algorithm depends on the hardware systems. We use the MATLAB (R2018a) in the CPU with 16 GB of memory and 3.40 GHz processor.

C. *Effect of HOG parameters*

In this section, we evaluate classification performance of the proposed scheme by varying the parameters of the HOG feature, i.e., cell size (s) and orientation bin size (n) of the histogram. The two parameters are the hyperparameters to use the HOG feature. It means that normally the optimal parameters are depend on the application datasets and should be determined empirically. In our experiment, using the grid search manner, we first optimize the cell size (s) while fixing the same bin size (n) as 9 which is optimized for pedestrian detection in the original HOG paper [25]. Then, using the selected optimal cell size (s), we vary the bin size (n).

Table 3 lists the comparison results of classification performance. For this comparison, we set the same dictionary size 70 and the threshold value 3 so that our focus will be only on the effect of HOG parameters.

First five rows in Table 3 list varying cell size (s) with the same bin size of 9, where we can note that $s = 64$ with 9 bins has the best result. In initial HOG paper [25], the cell size of 8 with 9 bins had the best classification performance for the specific human detection application. Since size of human subjects is very similar and subjects are always upright in images [25], smaller cell size could create more effective local features for classification. However, in this study, larger cell size ($s = 64$ or 128) shows better performance than smaller one. It might be because in this study, polyp position and size are not aligned. In addition, we already extract small patch image including polyp part in the patch extraction step. Therefore, smaller cell size for extracting local feature in the HOG method causes difficulty to capture whole polyp characteristics using the extracted patch image in our framework. This might be the reason for degraded classification performance with small cell size in Table 3.

For the bin size (n), i.e., the last 4 rows and the fourth row in Table 3, the bin size $n = 9$ or 12 with $s = 64$ shows improved performance than smaller bin size. The accuracy difference between the bin size 9 and 12 is very small (0.0027) and negligible. However, if we choose too small bin size, e.g., n=3 in Table 3, it may not be able to represent enough orientation information and results in accuracy decrease. On the other hand, too large bin size, e.g., n=15, may lead to overfitting problem.

D. *Effect of dictionary size*

In the dictionary learning scheme explained in Section II-*D*, dictionary size, i.e., the number of columns in the dictionary, should be initially determined to solve equation (2). It depends on the size of feature dimension, the number of classes, and application data. Normally, it is determined empirically [23].

Figure 6 shows the classification results by varying the dictionary size in the proposed scheme. Here, we compare classification accuracy for two different setups of HOG parameters ($s = 64$, $n = 9$ and $s = 64$, $n = 12$). We note that too small number of atoms in the dictionary show poor classification accuracy. It because if the number of bases (atoms) in dictionary is too small, new test feature cannot be represented well with similar features in the dictionary. Therefore, it resulted in low discrimination power. Figure 6 from 60 to 80 shows good classification performance for both HOG parameter cases (red and blue lines). On the other hand, if the number of atoms is too large, it is possible to include undistinguishable training features for both classes in the dictionary. In addition, computation time is also increased with large dictionary size.

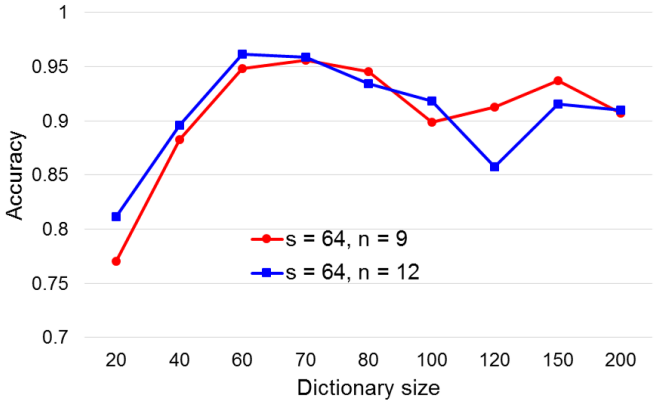

Figure 6. Comparison of classification accuracy for different dictionary size

E. *Effect of threshold value in polyp classification*

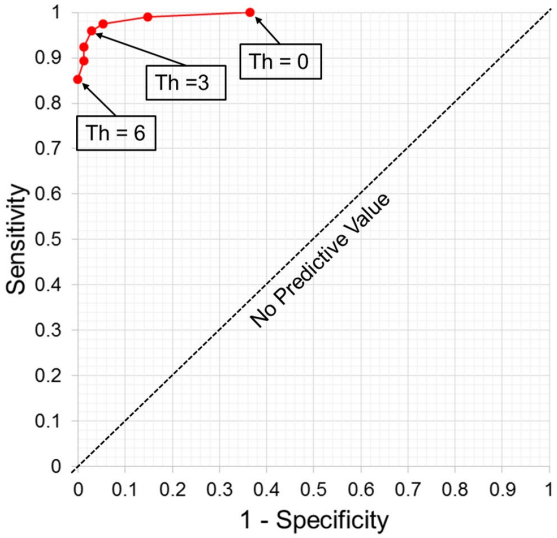

Figure 7. ROC curve analysis for threshold value

As we mentioned in Section II-*E*, we introduce a threshold value (Th) for the purpose of whole image classification. In this subsection, we examine the effect of the threshold value. Figure 7 indicates the ROC (receiver operating characteristic) curve when the threshold value is varied between 0 and 6. ROC curve represents both sensitivity and specificity simultaneously when its discrimination threshold is varied. It is well known as an effective tool for evaluation of diagnostic tests [43]. Figure 7 shows that the optimal threshold value is 3 for a balanced

performance of both sensitivity (0.9592) and specificity (0.9588). Here, we use HOG parameters s = 64, n = 12 and dictionary size of 70.

Note that in some medical diagnostic applications such as polyp or cancer detection, higher sensitivity with reasonable specificity is more important. This is because finding abnormal parts is the main goal of a clinical diagnostic system. Therefore, for obtaining higher sensitivity, i.e., correctly classified polyp images over total number of actual polyp images, a bit decreased specificity can be allowed. This decreased specificity leads to miss-classification case that some normal images having undistinguishable polyp parts are classified as polyp images. However, within the computer-aided polyp classification procedure (shown in Figure 1) this miss classification case might be acceptable although clinician will spend some time for searching to find those difficult-to-find polyps.

As shown in Figure 7, we can control the trade-off relationship between sensitivity and specificity by changing the threshold value (Th) in the proposed scheme. For example, we can set the threshold value to 2 (see Figure 7). This means the sensitivity is improved (0.9796) with a bit decreased specificity value (0.9176). This control function might be helpful for clinical diagnostic applications in practice.

F. Comparison with CNN based approach

The CNN-based deep learning approach is considered as state-of-the-art technique in many image recognition applications including polyp detection [14][44][47]. In this section, we aim to compare proposed method with the CNN-based deep learning approach by applying the same image patch based polyp screening framework.

CNN learns features from raw image pixels without any information about the images. Essential steps for CNN are convolution and pooling layer. In the convolution layer, spatial convolution between predefined filter and pixel values in a local region is performed. In the pooling layer, subsampling is performed for summarizing feature responses around neighboring pixels. Pooling operation makes learned features spatially invariant to the location of object images. Usually max pooling operation is widely used in CNN based image classification [14][47]. For CNN-based polyp screening, we use the same image preprocessing and patch extraction methods which are mentioned in Section II-A. Then, the 3-channel (RGB) based raw image pixels are used for training of CNN for the purpose of patch image classification.

Figure 8 shows the CNN architecture used in this study. In the CNN architecture, there are tunable hyperparameters such as the number of convolution filters, size of convolution filter and size of pooling area. In this study, we performed exhaustive gird search for optimization of CNN hyperparameters. Furthermore, we also optimize the number of convolution-pooling layers by varying from shallow to deep. Three convolution-max pooling layers are adopted followed by a fully connected layer with 256 nodes. We use 32 convolution filters for each convolution layer with 5×5, 3×3 and 3×3 filter size. For the max pooling operation, 3×3, 2×2 and 2×2 pooling area is chosen for each layer. The Adamax optimizer is used for CNN optimization with a default learning rate of 0.002 [46].

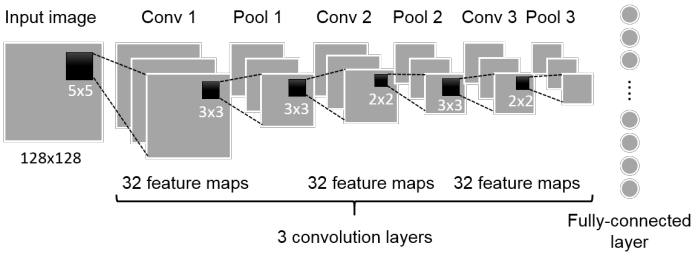

Figure 8. Three convolution-pooing layer based CNN architecture

After the image patch-classification using CNN framework, whole image frame-classification is performed by the same threshold based method introduced in Section II-*E* is performed. The number of epochs to train is set to 100.

Table 4 lists the comparison results between the proposed method and the CNN-based approach. Proposed combined feature based dictionary learning framework shows better classification performance, i.e., sensitivity, specificity, accuracy and precision, than the CNN-based deep learning method.

TABLE 4. COMPARISON OF CLASSIFICATION PERFORMANCE BETWEEN PROPOSED METHOD AND CNN-BASED DEEP LEARNING METHOD

Methods	Sensitivity	Specificity	Accuracy	Precision
Proposed method	0.9592	0.9588	0.9590	0.9641
CNN-based deep learning	0.9184	0.9235	0.9208	0.9326

Note that normally huge training dataset is used for training of CNN in image classification applications [46][47]. However, in the medical domain, annotated public dataset is limited and not always available. In this case, the proposed hand-craft feature based dictionary learning scheme might be a good machine learning tool for computer-aided detection.

G. Comparison with polyp localization studies

As mentioned in Section I, finding the exact location of polyps in an image frame is a very challenging task. In 2015 there was a competition for automatic polyp localization in colonoscopy [44]. This grand challenge was conducted at the 2015 International conference on Medical Image Computing and Computer Assisted Intervention (MICCA) [42]. For this competition on frame analysis, they used the same dataset like us; CVC-Clinic dataset for the training stage whereas ETIS-Larib for testing stage.

The results summarized in [44] were for seven teams where they showed the best classification performance was obtained using the CNN-based approach. Their results were 72.3% precision and 69.2% sensitivity (recall). This means with this performance on localization, it cannot be directly used as a clinical colonoscopy diagnostic tool. Because our study is on the screening of polyp image frame, the performance of the proposed method in this paper cannot be directly compared with

those in [44]. However, finding polyps in early stage is very important to prevent CRC. We expect that with high classification performance the proposed automatic polyp screening method can be useful for clinical colonoscopy diagnostic tool to reduce polyp miss-detection rate and enhance the clinician's performance.

IV. Conclusion

In this paper, we propose a dictionary based learning scheme for automatic polyp screening in colonoscopy image data. Small patch image is extracted from the whole image frame by using a sliding window method in combination with a combined color and shape feature to effectively capture polyp characteristics. Furthermore, combined feature based dictionary is obtained by the dictionary learning procedure, where the dictionary is used to extract the final feature vector using the sparse coding step. A linear SVM is applied for patch-image classification and a simple thresholding performs the final whole-image classification. The proposed polyp screening framework is evaluated using three public colonoscopy datasets. Our experimental results show that the proposed combined feature based dictionary learning scheme outperforms individual shape and color feature based method and also combined feature based method without dictionary learning procedure. Furthermore, the proposed hand-craft feature based dictionary learning scheme shows better classification performance than the CNN based deep learning approach within the same patch image based polyp screening framework.

Acknowledgment

The authors would like to express sincere appreciation to Dr. Ali Khaleghi and Dr. Ali Chelli at Norwegian University of Science and Technology for their valuable comments.

References

[1] R. L. Siegel, K. D. Miller and A. Jemal, "Cancer statistics 2016," *CA Cancer J Clin.,* vol. 66, pp. 7-30, 2016.

[2] Y. Yuan, B. Li and M. H. Meng, "Improved bag of feature for automatic polyp detection in wireless capsule endoscopy images," *IEEE Trans. Autom. Sci. Eng.,* vol. 13, no. 2, pp. 529–535, 2016.

[3] D. Lieberman, "Quality and colonoscopy; a new imperative," *Gastrointest Endosc.*, vol. 61, pp. 392-394, 2005.

[4] A. M. Leufkens, M. G. H. van Oijen, F. P. Vleggaar and P. D. Siersema, "Factors influencing the miss rate of polyps in a back-to-back colonoscopy study," *Endoscopy*, vol.44, no. 5, pp. 470–475, 2012.

[5] L. Rabeneck, J. Souchek and H. B. El-Serag, "Survival of colorectal cancer patients hospitalized in the Veterans Affairs Health Care System," *Am J Gastroenterol.*, vol. 98, no. 5, pp. 1186-1192, 2003.

[6] T. Ma, Y. Zou, Z. Xiang, L. Li, and Y. Li, "Wireless capsule endoscopy image classification based on vector sparse coding," *in Proc. IEEE China Summit Int. Conf. IEEE Signal Inf. Process*., 2014, pp. 582–586.

[7] S. A. Karkanis, D. K. Iakovidis, D. E. Maroulis, D. A. Karras, and M. Tzivras, "Computer-aided tumor detection in endoscopic video using color wavelet features," *IEEE Trans. Inf. Technol. Biomed.*, vol. 7, no. 3, pp. 141–152, Sep. 2003.

[8] L. A. Alexandre, N. Nobre, and J. Casteleiro, "Color and position versus texture features for endoscopic polyp detection," *in Proc. IEEE Int. Conf. BioMed. Eng. Informat.*, 2008, vol. 2, pp. 38–42.

[9] S. Hwang, J. Oh, W. Tavanapong, J. Wong, and P. de Groen, "Polyp detection in colonoscopy video using elliptical shape feature," *in Proc. IEEE Int. Conf. Image Process.*, 2007, vol. 2, pp. II-465–II-468.

[10] J. Bernal, J. Snchez, and F. Vilario, "Towards automatic polyp detection with a polyp appearance model," *Pattern Recognit.*, vol. 45, no. 9, pp. 3166–3182, 2012.

[11] J. Bernal, J. Snchez, G. F.-Esparrach, D. Gil, C. Rodriguez and F. Vilario, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Comput. Med. Imag. Graph.*, vol. 43, pp. 99–111, 2015.

[12] S. Bae and K. Yoon, "Polyp detection via imbalanced learning and discriminative feature learning," *IEEE Trans. Med. Imag.*, vol. 34, no. 11, pp. 2379–2393, Nov. 2015.

[13] N. Tajbakhsh, S. R. Gurudu and J. Liang, "Automated Polyp Detection in Colonoscopy Videos Using Shape and Context Information," *IEEE Trans. Med. Imag.*, vol. 35, no.2, pp. 630-644, 2016.

[14] S. Park, M. Lee, and N. Kwak, "Polyp detection in colonoscopy videos using deeply-learned hierarchical features," *Seoul Nat. Univ.*, 2015.

[15] S. Park and D. Sargent, "Colonoscopic polyp detection using convolutional neural networks," *SPIE Med. Imag.*, p. 978528, 2016.

[16] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3736-3745, Dec. 2006.

[17] J. Yang, J. Wright, T. Huang and Y. Ma, "Image super-resolution via sparse representation," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2861-2873, Nov. 2010.

[18] R. G. Baraniuk, "Compressive sensing," *IEEE Signal Processing Mag.*, vol. 24, no. 4, pp. 118–120, 124, July 2007.

[19] J. Wright, A. Yang, A. Ganesh, S. Sastry and Y. Ma, "Robust face recognition via sparse representation", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 210-227, Feb. 2009.

[20] S. Mallat, A Wavelet Tour of Signal Processing, 1998, Academic.

[21] M. Aharon, M. Elad and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311-4322, Nov. 2006.

[22] R. Raina, A. Battle, H. Lee, B. Packer and A.Y. Ng, "Self-Taught Learning: Transfer Learning from Unlabeled Data," *Proc. 24th Int'l Conf. Machine Learning*, 2007, pp. 759-766.

[23] J. Mairal, F. Bach, J. Ponce and G. Sapiro, "Online dictionary learning for sparse coding", *ICML*, 2009.

[24] M. Liu, "Sparse classification for computer aided diagnosis using learned dictionaries", *Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2011, vol. 6893, pp. 41-48.

[25] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *CVPR*, pages I, pp. 886-893, 2005.

[26] O. Deniza, G. Buenoa and J. Salido, "Face Recognition Using Histograms of Oriented Gradients," *Pattern Recognition Letters*, vol. 32, no. 12, pp. 1598-1603, 2011.

[27] Y. Iwahori, A. Hattori, Y. Adachi, M. K. Bhuyan, R. J. Woodham and K. Kasugai, "Automatic Detection of Polyp Using Hessian Filter and HOG Features," *Procedia Computer Science*, vol. 60, pp. 730-739, 2015.

[28] P. Pujas and M. Aldon, "Robust colour image segmentation", *In 7th International Conference on Advanced Robotics (ICAR'95)*, 1995.

[29] P. Paclik, J. Novovicova, P. Somol and P. Pudil, "Road sign classification using Laplace kernel classifier", *Pattern Recognition Lett.*, vol. 21, no. 13-14, pp. 1165-1173, 2000.

[30] W.T. Wang, C. Xu and H.W. Shen, "Eye localization based on hue image processing", *International Symposium on Intelligent Signal Processing and Communication Systems*, pp. 730-733, 2007.

[31] H.F. Hashem, "Adaptive technique for human face detection using HSV color space and neural networks", *Radio Sci. Conf. NRSC 2009*, pp. 1-7, 2009.

[32] F. J. C. Condessa, "Detection and Classification of Human Colorectal Polyps," IST, Lisbon, Portugal, 2011.

[33] Y. Chen and J. Lee, "A review of machine-vision-based analysis of wireless capsule endoscopy video", *Diagnostic and Therapeutic Endoscopy*, 2012.

[34] T. Ghosh, S. K. Bashar, S. A. Fattah, C. Shahnaz and K. A. Wahid, "An automatic bleeding detection scheme in wireless capsule endoscopy based on statistical features in hue space", *17th International Conference on Computer and Information Technology, ICCIT 2014*, pp. 354-357.

[35] T. Gevers and A. W. M. Smeulders, "Color based object recognition", *Pattern Recognit.*, vol. 32, pp. 453-465, Mar. 1999.

[36] A. Y. Yang, Z. Zhou, A. G. Balasubramanian, S. S. Sastry and Y. Ma, " Fast $l1$-minimization algorithms for robust face recognition ", *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3234-3246, Aug. 2013.

[37] D. Donoho and Y. Tsaig. “Fast solution of l1-norm minimization problems when the solution may be sparse,” preprint, 2006. http://www.stanford.edu/ tsaig/research.html
[38] M. Asif and J Romberg L1 Homotopy (http://users.ece.gatech.edu/~sasif/homotopy/)
[39] S. Ameling, S. Wirth, D. Paulus, G. Lacey and F. Vilario, "Texture-based polyp detection in colonoscopy" in Bildverarbeitung fr die Medizin 2009, pp. 346-350, 2009, Springer.
[40] MathWorks: http://www.mathworks.co.kr/kr/help/stats/support-vector-machines-svm.html
[41] J. S. Silva, A. Histace, O. Romain, X. Dray and B. Granado, “Towards embedded detection of polyps in WCE images for early diagnosis of colorectal cancer,” *Int J Comput Assist Radiol Surg.*, vol. 9, no. 2, pp. 283-293, 2014.
[42] Automatic polyp detection in colonoscopy videos: https://grand-challenge.org/site/polyp/databases/
[43] K. Hajian-Tilaki, "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation", *Casp. J. Intern. Med.*, vol. 4, no. 2, pp. 627-35, Jan. 2013.
[44] J. Bernal, N. Tajkbaksh,, F. J. Sánchez, J. Matuszewski, H. Chen, L. Yu, Q. Angermann, O. Romain, B. Rustad, I. Balasingham, K. Pogorelov, S. Choi, Q. Debard, L. M. Hen, S. Speidel, D. Stoyanov, P. Brandao, H. Cordova, C. S. Montes, S. R. Gurudu, G. F. Esparrach, X. Dray, J. Liang and A. Histace, "Comparative Validation of Polyp Detection Methods in Video Colonoscopy: Results from the MICCAI 2015 Endoscopic Vision Challenge," *IEEE Trans. Med. Imaging*, vol. 36, no. 6, pp. 1231-1249, 2017.
[45] Y. Wang, W. Tavanapong, J. Wong, J. Oh and P. C. de Groen, "Part-Based Multiderivative Edge Cross-Sectional Profiles for Polyp Detection in Colonoscopy," *IEEE J Biomed Health Inform,* vol. 18, no. 4, pp. 1379-1389, July 2014.
[46] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *CoRR*, vol. abs/1412.6980, 2014.
[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., pp. 1097–1105, 2012.
[48] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway and Jianming Liang, “Convolutional neural networks for medical image analysis: Full training or fine tuning?” *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1299–1312, May 2016.
[49] Y. Shin, H. N. Lee and I. Balasingham, “Fast L1-based sparse representation of EEG for motor imagery signal classification”, *in Engineering in Medicine and Biology Society (EMBC), 38th Annual International Conference of the IEEE*, pp. 223–226, 2016.