Computer Aided Diagnosis Scheme for Polyp Detection in CT Colonography Using K-Means Clustering and SVM

K. Gayathri Devi, R. Radhakrishnan and Kumar Rajamani

1Anna University, Chennai ECE department, NGP Institute, Coimbatore, India
2Principal, Sri Sakthi Institute of Engineering and Technology, Coimbatore, India
3Research Engineer, GE Global Research, Bangalore, India

Abstract: CAD-CTC helps the radiologists for the automatic detection of suspicious polyps in the abdominal CT images of the colon. In this paper, a novel method is proposed in order to reduce the computational time taken by the CAD system for the detection of growth in abdominal CT images for the identification of colorectal cancer. In the proposed CAD system, segmentation of the colon is done by Otsu’s method of thresholding and clustered by K-Means clustering for the extraction of the features such as intensity texture and volume in order to classify the polyps. The segmented candidates are typically characterized by features describing, for instance, the polyp shape and its internal intensity distribution. Such features will serve as input for the classification system. Classification is performed by SVM (Support Vector Machine) and ANN (Artificial Neural Network). The proposed CAD system is evaluated with 16 slices (each) of normal and abnormal images. The CAD system achieved a minimal computation time (10 minutes) and an learning rate of about 2.62 at an epoch 23. Thus the good results demonstrate that the proposed system may provide relevant additional information for the clinical decision process.

Key words: CAD - CTC - Polyp - Support Vector machine (SVM)

INTRODUCTION

Colorectal cancer is cancer that occurs either in the colon or rectum. Most colorectal cancers develop first as colorectal polyps, which are growths inside the colon or rectum that may later become cancerous as shown in fig 1. It affects both men and women mostly above the age of 50. It is the third most common cancer among men after prostate and lung cancers. For women, colorectal cancer is the third most common cancer after breast and lung cancers [1].

The Colorectal Polyps can be classified as shown in the table 1. The two most common types are adenomas and hyperplastic polyps. Benign are tumors which rarely or never give rise to cancer but the cells in tubular adenomas frequently progress to cancer, show certain abnormalities of cell maturation and appearance collectively known as dysplasia.

The most important colorectal polyp is the adenoma, a small benign tumor growing to about 2 cm in size. Colonic adenomas are common and in the majority of patients there is no side effect on health. They are more common with increasing age. Appropriate evidence are shown that colonic adenomas are the early stage of colorectal cancer, although only a very small percentage of adenomas turn into malignant tumour and it may take nearly five to 15 years to change. The larger the polyp, the greater the probability that the polyp will have undergone malignant change and contain cancer (adenocarcinoma).

Fig. 1: Polyps situated on the colon wall

Corresponding Author: K. Gayathri Devi, Anna University, Chennai ECE department, NGP Institute, Coimbatore, India.
Table 1: Classification of colorectal polyps

<table>
<thead>
<tr>
<th>Histological Classification</th>
<th>Polyp Type</th>
<th>Malignant Potential</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-neoplastic</td>
<td>Hyperplastic polyps</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Hamartomas</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Lymphoid aggregates</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Inflammatory polyps</td>
<td></td>
</tr>
<tr>
<td>Neoplastic (adenomas)</td>
<td>Tubular adenomas (0–25% villous tissue)</td>
<td>Yes</td>
</tr>
<tr>
<td></td>
<td>Tubulovillous adenomas (25–75% villous tissue)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Villous adenoma (75–100% villous tissue)</td>
<td></td>
</tr>
</tbody>
</table>

Hyperplastic (or metaplastic) polyps are usually small, pale curved elevations of the colon lining. These are very common. Although hyperplastic polyps themselves do not turn into colorectal cancer, occasionally hyperplastic polyps (particular those which are large and multiple) will contain adenomas, known as mixed hyperplastic adenomatous polyps. In these polyps, development of cancer may occur but it is very rare.

Polyps are small growths that arise from the inner lining of the colon. CAD detects the polyps by identifying their characteristic "bump-like" shape fig 2.

Computed tomographic colonography (CTC) or virtual colonoscopy is an evolving method for detecting colon polyps which records the 3D image of the patient’s abdomen [2]. Virtual colonoscopy combines axial spiral CT data acquisition of the air-filled and cleansed colon with 3-dimensional imaging software to create endoscopic images of the colonic surface [3]. Two main factors that limit CT colonography are its excessive interpretation time and the variable sensitivity among readers. This paper focuses the advantages of computer-aided detection (CAD) techniques to overcome the above problems faced by the radiologist [4].

With the introduction of the multi slice CT scanners, the amount of data to be examined by the radiologists has increased dramatically and a dataset can consist of 800 to 2000 axial images depending on the patient height and the resolution of the CT data.

Thus, the visual examination of such a large amount of data is a time consuming procedure and the results are not always reproducible since they are often affected by eye fatigue and the subjectivity of the human operator. This has led to a sustained effort to develop computer-aided diagnosis (CAD) schemes that can help radiologists with locating colorectal lesions in CTC data in an efficient and accurate manner [5].

CAD is fundamentally based on highly complex pattern recognition. In recent years a number of prototype Colon CAD schemes with CTC have been proposed for detecting polyps to reduce the false positive (FP) polyps [6]. About 95% of all cases of colon cancers arise from adenomatous polyps are initially benign, but at later stage they turn into malignant. The malignancy is normally referred with the size of polyps. The size of a detected polyp is an important criterion for further diagnosing and decision making [7].

**Different CAD Schemes:** Polyps were classified as small (= 5 mm), medium (6 – 9 mm) and large (=10 mm) [8].

Wei Hong, et al., (2006) [9] have proposed a method which was implemented by first segmenting and digital cleansing the colon acquired using CT from abdominal data set by partial volume segmentation algorithm. This algorithm separates voxels in the air lumen into three categories and this step produces segmentation of the colon and a clean colon lumen. A region growing based segmentation algorithm is used to extract a topologically simple segmentation of the colon lumen. A region growing based segmentation algorithm is used to extract a topologically simple segmentation of the colon lumen. After this step, both the colon surface mesh with genus zero and the colon centerline is obtained. The 3D colon surface is mapped to a 2D rectangle and this conformal mapping simplifies our polyp detection problem from 3D to 2D. High resolution 2D biopsy images are generated using volumetric ray-casting algorithm by passing a ray which is allowed to traverse up to 40 steps. The K-L transformation matrix is applied to the local vector series belonging to hand segmented polyps on the 2D flattened electronic biopsy images. The representative vector V and similarity threshold T is used to classify the feature vectors in the K-L domain. If the Euclidean distance between a feature vector V is less than T, the corresponding pixel is classified as belonging to a polyp. A 2D image...
by implementing the wrapper method for feature selection, Tarik A. Chowdhury extraction phase in the classical polyp detection scheme greater than 50 mm it is to be considered as a larger polyp.

June-Goo Lee et al., (2011) [11] proposed a method for the identification of polyps (3-6 mm) by producing a numerical phantom. The filtering procedure begins with 3-D Gaussian filtering, followed by a Hessian matrix computation and its eigen decomposition. This phantom is used for the evaluation of effectiveness of the volumetric structural features such as Gaussian kernel (GK), suppression threshold (ST) and volume threshold (VT) to discriminate basic structures in the colon. The former two are parameters of the blob-likeness filter and the third was used for candidate detection.

Vincent F. et al., (2010) [12] proposed a system for the identification of polyps. This system was implemented by using candidate extraction and a logistic classifier. The extracted candidates from the candidate extraction stage are ordered by a linear logistic classifier which classifies based on only three features: the protrusion of the colon wall, the mean internal intensity and a feature to discard detections on the rectal enema tube.

Simona E. Grigorescu, et al., (2010) [13] have proposed a method for the automatic detection and segmentation of polyps in diameter. This paper modifies the segmentation phase in the classical polyp detection scheme by evaluating the LH histogram which separates the fat and muscle density in CTC data. It focuses mainly on the tissue separation surrounding the colon lumen and finds the appropriate region for poly detection. This Large lesion detection is started with a preselection step, which is divided into many stages. The first stage removes the image noise and the small food remnants by using the intensity information. The next stage in the preselection step is to group the location that passed the above stage based on the proximity of closeness to each other. The segmentation is done based on the LH histogram representation of the local gray value data. LH histogram is a method to identify the boundaries where L and H represent the lowest and highest intensity value. The points that belong to the material and border are plotted in the LH space, for materials centered at (L, L) and for borders centered at (L, H). If two materials are to be segmented LH space is split into four quadrant and by repeated iteration suitable threshold that separates the two material can be found out. After segmentation and performing morphological operation to remove the small connection between muscles like structure, the diameter of the segmented region is calculated if it has a diameter greater than 50 mm it is to be considered as a larger polyp.

Tarik A. Chowdhury et al. (2008) [14] have proposed a method for the detection and segmentation of polyps when applied to low dose real patient data. The important phase in the polyp detection, automatic colon segmentation was performed based on the normalized h-gram. This similar type of colon segmentation was also proposed by [15]. The candidate generation is performed with Surface Convexity Test, which evaluates the convexity of all voxels that define the candidate surface. The features that were used for the identification of polyps were the standard deviation (SD) of the surface variation, the SD of the three axes of the fitted ellipsoid, SD of the ellipsoid fit error and SD of the sphere fit error and the value of the Gaussian distribution. The classifications of polyps were performed by feature normalized nearest neighbour (FNNN) classifier and the probabilistic neural network (PNN) classifier.

David S. Paik et al. (2004) [16] proposed the techniques as follows. The Preprocessing and Segmentation of colon lumen is done by thresholding all the air intensity voxels < -700 HU including air outside the body. The inferior portion of lungs is removed by using negative masking 3-D region-filling seeded with air intensity regions with a width or depth of greater than 60 mm. The binary image S2 is derived which is limited only to the edges of air-tissue interfaces. First derivative gradient operation, a modification of the canny edge detector is applied to binary image S2 to identify the edges in S2 resulting in 3-D orientation of the image surface normals represented as N(x,y,z). The surface normal overlap step is next performed for detecting lesions. In addition to this, theoretical model were also developed to compare the behaviour of the surface normal overlap (SNO) method to the 3-D Hough transform for spheres to distinguish colonic polyps from folds.

In the consensus [17] proposed that patients with a polyp of at least 10 mm must be referred to optical colonoscopy for polypectomy and it is advised that diminutive polyps (5 mm) should not even be reported but still detection of small and diminutive are performed [18].
Marcelo Fiori et al. (2011) [19] have proposed CAD system. This paper segment the colon from the abdominal volume in the CT by proposing a probability map for three classes of interest air, liquid and interface, the union of these three classes P forms the inside of the colon. The segmentation of air and liquid can be done manually based on the distribution of the gray level value. The segmentation of interface is done by a predefined algorithm that computes the probability of that interface. Thus the union of these three classes P segments the colon from the CT slice. The curvature-based surface motions with suitable modification is implemented to smoothen the colon surface and to remove the noise after segmentation step. Feature Extraction extracts 6 geometric features by considering the shape of neighborhood area of the polyp in addition to polyp intrinsic geometry and structure. In addition to feature extraction 5 texture feature entropy, energy, contrast, sum Mean and homogeneity are also extracted to aid the classification and extraction of polyps.

Anna K. Jerebko et al. (2003) [20] implemented the CAD technique in the two-dimensional subimages of polyp candidates selected using various 3-D shape and curvature characteristics. Preprocessing was done with median filter to remove salt and pepper noise. Edges are the boundaries of the colon wall with the surrounding tissue. Among the various edge operators which detect the edges canny edge detector is used to identify the polyp boundaries. The proposed algorithm is able to quantitatively measure particular features (polyp boundary length, number of boundary pixels, polyp base length, polyp internal area, mean intensity, polyp height and inscribed circle radius) of the polyp and use them for classification. After applying the canny filter, the Radon transformation is applied to the edge detected image. The Radon transform computes projections in the contour image c(x,y) along specified directions. This projection parameters help in extracting and measuring polyp features.

The Rest of the Section Is Organized as Follows: Section II briefs the overview of the proposed work, Section III will narrate the steps for segmentation and extraction of candidate in the CT images. Extraction of features such as intensity texture and volume will be addressed in section IV. This section will be followed by the next phase Classification and Evaluation and the conclusion and validation will be the last section.

Overview of the Proposed System: The classical CAD system for the detection of polyps can be generally decomposed into four stages [8]. The four phases are segmentation, candidate extraction, classification and evaluation as shown in figure 3. The automatic method of extracting colon is itself a challenging task independent of the scanner devices and settings. There are papers which focuses entirely on the concentration of colon also [15]. In the proposed CAD system, segmentation of the colon is performed by otsu’s method of thresholding and clustered by k-means for extraction of the candidates. Classification is performed by SVM (Support Vector Machine) [3] and ANN (Artificial Neural Network) from the training and testing images of candidate extraction. The proposed CAD system is evaluated with 16 slices (each) of normal and abnormal images. The CAD system achieved a minimal computation time (10 minutes) and an learning rate of about 2.62 exp-006 at an epoch 23. This novel method is proposed in order to decrease the computational time taken by the CAD system.

The above steps were implemented with both normal and abnormal dataset of Abdominal CT images of thickness 0.6mm. The proposed system were evaluated with the datasets from Clarity imaging centers.

Segmentation and Candidate Extraction: Segmentation is performed on the abdominal CT images by using the otsu’s method of thresholding [21]. Fig.2 shows the sample abdominal CT images obtained from the clarity imaging centre for the evaluation of this proposed method. The original images shown in Fig. 4 (a) and (b).
Algorithm

Step 1: Images are resized to 512x512

Step 2: The images are converted into gray scale after which otsu’s thresholding is performed.

Step 3: If the value of the pixel is zero then it returns black else it returns white.

Step 4: Then binary components are connected.

Step 5: The area which contains tumor is chosen from the properties of the image (800:70000). Then the image is contrast adjusted and the segmented colon is shown in Fig.5 (a) and (b). The candidate extracted colon is shown in Fig.6 (a) and (b).

Now the candidate extraction will be done from the segmented colon. K-means clustering [5] is used because it is simple and has relatively low computational complexity. In addition, it is suitable for biomedical image segmentation as the number of clusters is usually known for images of particular regions of human anatomy.
By the algorithm the extraction of the candidates is performed on the segmented part by means of K-Means Clustering.

Each data point is analyzed and shifted such that it is close to the cluster that is formed. If the data point is already closest to its own cluster, it will not be shifted. The process will continue until cluster means do not shift more than a given cut-off value or the iteration limit is reached.

### Feature Extraction:

Feature extraction is to identify and separate desired portions or shapes of a image. Colon lesion can be classified based on the parameters such as size, diameter of the growth, or according to their shapes, into pedunculated, sessile or flat. In this paper feature extraction phase the focus is on the extraction of three different parameters surface intensity characteristics, volumetric and surface shape characteristics and texture characteristics. The flow chart for feature extraction is shown in fig 6. where features such as volume, intensity and texture are determined [6].

In this study, straightforward approaches were presented for polyp detection in CTC. The underlying principles were a single polyp-specific volumetric feature could effectively distinguish polyps from the other colonic structures.

For the selected 16 images of both normal and abnormal, the features such as intensity texture and volume are determined and tabulated in Table 2.

The internal intensity of the candidates has been found to be a best feature to reduce a large number of false detections. The haustral fold in the colon may be misinterpreted as polyp or growth in the identification of cancer.

Texture is a powerful parameter that helps in the retrieval process. Texture on its own does not have the capability of finding similar images, but it can be used to classify textured images from non-textured ones and then be combined with another visual attribute like color to make the retrieval more effective.

### Classification and Evaluation:

Classification of CT image data aims at discriminating multiple objects from each other within the image. Classification will be executed on the base of spectral or spectrally defined features, such as texture, volume and intensity in the feature space. It can be said that classification divides the feature space into several classes based on a decision rule. Candidate uniquely identified with the above associated features are then fed as input to a classifier for final evaluation with the help of training set or test set. The classification and evaluation is performed by using two methods.

They are:

- Support Vector Machine (SVM) classification
- Artificial Neural Networks (ANN)

The numerical results were obtained by classifying with SVM using cost sensitive learning after normalizing the data as shown in fig 6. The goal of this classification is to separate normal images from abnormal images using three Features which have been extracted in the feature extraction phase. The normal and abnormal images were clearly separated as shown in figure 7.
The evaluation were performed with 32 abdominal slices of CT images (16 normal and 16 abnormal images) using ANN (artificial neural network) to evaluate performance. The neural networks were trained with back propagation method consisting of three layers input layer, 20 hidden layer and the output layer. The Fig. 8 reveals the graphical representation over the number of epochs that are needed for attaining the best performance evaluation and the Mean Square Error Rate (MSR) at the phase of neural network classification. The best validation performance were obtained as 2.62e-006 at 23 epochs with minimal MSE. Besides the related methods the classification using the above feature produced good performance rate with less mean square error (MSE) rate as shown in fig. 8.

**CONCLUSION**

Thus this method of screening technique of CAD-CTC (Computer Aided Diagnosis-Computer Tomography Colonography) with variations in the segmentation and classification has been really effective in identifying the exact polyps. Thus, this paper helps the radiologists to detect the exact polyp by distinguishing it
(polyps) from the folds and false positives. The existing system and the proposed system is compared by using the parameter as shown in the table. Thus the proposed system demonstrated good results such as the low computation time and high learning rate and also low mean square error rate as shown in the figure 9.

REFERENCES


