Transactions on Machine Learning
and Data Mining
Vol. 4, No. 2 (2011) 96-115

ISSN: 1865-6781,
ISBN: 978-3-942952-07-1

ibai Publishing
www.ibai-publishing.org

Mining Colorectal Polyp Images for Colon Examination based on Texture Description and Decision Tree Induction

Anja Attig and Petra Perner

Institute of Computer Vision and applied Computer Sciences, IBaI
PF 30 11 14, 04251 Leipzig, Germany
pperner@ibai-institut.de, www.ibai-institut.de

Abstract. Medical disease examination is often based on images. Mining these images in order to obtain the classification knowledge for automatic image classification is a challenging task. This task belongs to the field of image mining. Image mining is usually not only comprised of mining a table of numbers it has also to do with transforming the image in the right image description. Both, the image description and the classification knowledge, determine the quality of the classifier. Texture is a powerful method to describe the appearance of different biological objects in images. There are different texture descriptors around. Which one is the best for medical images is still an open question. The most used texture descriptor is the well-known Haralick´s texture descriptor. We propose a texture descriptor based on random sets. This descriptor gives us more freedom in describing different textures. In this paper we develop two classification models based on decision tree induction, one for each of the two texture descriptors. We compare the two texture descriptors based on a medical data set. We review the theory of the two texture descriptors and describe the procedure for the comparison of the two methods. A medical data set is used that is derived from colon examination. Decision tree induction is used to learn a classifier model. Cross-validation is used to calculate the error rate. The comparison of the two texture descriptors is based on the error rate, the properties of the two best classification models, the runtime for the feature calculation, the selected features, and the semantic meaning of the texture descriptors.

1 Introduction

Texture is a powerful method to describe the appearance of different biological objects in images. Patterns on cells in cell images, on fungi images or polyp images can be described by texture.

Different texture descriptors have been developed over the past [1]. The most used texture descriptor is the well-known Haralick´s texture descriptor [2]. Although it works well on different applications we prefer to use our texture descriptor [3] that is based on random sets [4] since this descriptor gives us more freedom in describing different textures.

In this paper we develop two classification models based on decision tree induction, one for each of the two texture descriptors. We compare the two texture descriptors based on a medical data set.

The theory of the two texture descriptors is reviewed in Section 2. The procedure for the comparison of the two methods is described in Section 3. The used data set of images is derived from colon examination. We calculated the texture features based on the two methods for each image of the data set and learn a decision tree classifier. Cross-validation is used to calculate the error rate. Then we compare the properties of the two best decision trees, the runtime for the feature calculation, the selected features, and the semantic meaning of the texture descriptors. The results are presented in Section 4. If feature preselection is improving the model and if it is a good preprocessing step is dissussed in Section 5. The method of calculating the number of class images for the texture descriptor based on Random Sets is disussed in Section 6. The discussion of the work is given in Section 7. Conclusions are presented in Section 8.

2 Texture Descriptors

The well-known Harlick´s texture descriptor [2] is the most used one among different texture descriptors. A texture descriptor based on random sets [4] is another texture descriptor that is flexible and powerful. In this section we review the theory of both texture descriptors.

2.1 Haralick´s Texture Descriptor

A co-occurrence matrix $C_{(\Delta x,\Delta y)}$ with the offset $(\Delta x, \Delta y)$ is defined over an *n*x*m* Image *I*:

$$C_{\Delta x,\Delta y}(i,j) = \sum_{p=1}^{n}\sum_{q=1}^{m}\begin{cases}1, & I(p,q)=i \text{ and } I(p+\Delta x, q+\Delta y)=j\\ 1, & I(p,q)=j \text{ and } I(p+\Delta x, q+\Delta y)=i\\ 0, & \text{otherwise}\end{cases} \quad (1)$$

The co-occurrence matrix can be interpreted as a matrix of frequency from neighboring pixels in image *I* with an offset $(\Delta x, \Delta y)$ where a pixel has the gray level *i* and the other pixel a gray level *j*. Note that this matrix is symmetric.

Let

$$P_{(\Delta x,\Delta y)} = \frac{1}{R} C_{(\Delta x,\Delta y)} \tag{2}$$

be the normalized co-occurrence matrix from $C_{(\Delta x,\Delta y)}$ with $R = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} C_{(\Delta x,\Delta y)}(i, j)$ being the normalized factor.

Let p_{ij} be the (*i*, *j*)-th element of matrix $P_{(\Delta x,\Delta y)}$, with N_g being the number of distinct gray levels in the image *I*. The *i-th* entry in the marginal probability matrix obtained by summing the rows of p_{ij} is

$$p_x(i) = \sum_{j=1}^{N_g} p_{ij} \quad \text{and} \quad p_y(i) = \sum_{i=1}^{N_g} p_{ij} \quad \text{for the lines, respectively.} \tag{3}$$

Further we are calculating:

$$p_{x+y}(k) = \sum_{\substack{i=1 \\ i+j=k}}^{N_g} \sum_{j=1}^{N_g} p_{ij} \quad \text{with} \quad k = 2, 3, \ldots, 2N_g \tag{4}$$

and

$$p_{x-y}(k) = \sum_{\substack{i=1 \\ |i-j|=k}}^{N_g} \sum_{j=1}^{N_g} p_{ij} \quad \text{with} \quad k = 0, 1, \ldots, N_g - 1. \tag{5}$$

since $P_{(\Delta x,\Delta y)}$ is also a symmetric matrix $p(i) = p_x(i) = p_y(i)$.

This results in:

$$\mu = \mu_x = \mu_y = \sum_{k=1}^{N_g} k p(k) \tag{6}$$

$$\sigma^2 = \sum_{k=1}^{N_g} p(k)(k - \mu)^2 \tag{7}$$

From that Haralick et. al [2] derived 13 features that are given in table 1.

Table 1. Haralick Texture Features

1. Angular Second Moment $f_1 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_{ij}^2$	2. Contrast $f_2 = \sum_{k=0}^{N_g} k^2 \left(\sum_{\substack{i=1 \\ \lvert i-j\rvert=k}}^{N_g}\sum_{j=1}^{N_g} p_{ij} \right)$
3. Correlation $f_3 = \frac{1}{\sigma_x \sigma_y}\left(\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} ijp_{ij} - \mu_x \mu_y \right) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} ijp_{ij} - \mu^2 \right)$	4. Sum of Squares: Variance $f_4 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} (i-\mu)^2 \, p_{ij}$
5. Inverse Difference Moment $f_5 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{1}{1+(i-j)^2} p_{ij}$	6. Sum Average $f_6 = \sum_{k=2}^{2N_g} k p_{x+y}(k)$
7. Sum Variance $f_7 = \sum_{k=2}^{N_g} (i - f_6)^2 \, p_{x+y}(k)$	8. Sum Entropy $f_8 = -\sum_{k=2}^{2N_g} p_{x+y}(k) \log p_{x+y}(k)$
9. Entropy $f_9 = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_{ij} \log p_{ij}$	10. Difference Variance $f_{10} = \text{variance of } p_{x-y}$
11. Difference Entropy $f_{11} = -\sum_{k=0}^{N_g-1} p_{x-y}(k) \log p_{x-y}(k)$	12. and 13. Information Measure of Correlation $f_{12} = \frac{f_9 - HXY1}{H}$ $f_{13} = \sqrt{1-\exp[-2(HXY2 - f_9)]}$
with $HXY1 = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_{ij} \log\left(p_x(i)p_y(j)\right)$, $HXY2 = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_x(i)p_y(j) \log\left(p_x(i)p_y(j)\right)$ and $H = \sum_{k=1}^{N_g} p(k) \log p(k)$	

This results in thirteen texture features.

For our tests we compute four co-occurrence matrixes $C_{(\Delta x, \Delta y)}$ with the offsets shown in Table 2.

Table 2. Used Offsets $(\Delta x, \Delta y)$ and their Neighborhood Relations

Angle	$(\Delta x, \Delta y)$
0°	(1,0)
45°	(-1,-1)
90°	(0,-1)
135°	(1,-1)

First we compute for the four matrixes the 13 Haralick Texture Features shown in table 1. Thus, we obtain four values for each feature. In order to reduce the feature set, we compute anolog to [2] the mean and rank of each feature so that we finally get 26 features. This feature set is named Haralick-1.

Another method to compute the texture features that saves computation time is to sum over the four matrixes CT

$$CT_{ij} = C_{(0,1)}(i,j) + C_{(-1,1)}(i,j) + C_{(-1,0)}(i,j) + C_{(-1,-1)}(i,j) \quad \textbf{(8)}$$

The matrix CT is normalized according to formula 2. From the normalized matrix PT we calculate the 13 texture features of Haralick`s descriptor. This feature set is named Haralick-2.

The feature set named Haralick-3 takes from the four matrixes all 13 features. The total number of features in this data set is fifty two.

In addition to the above described features, Haralick defines the Maximal Correlation Coefficient as feature number 14; that we do are not using this coefficient in our study because of the high computation time.

2.2 Texture Descriptor Based on Random Sets

Our texture descriptor is based on Boolean sets. Boolean sets were introduced by Matheron [5]. An in-depth description of the theory can be found in Stoyan et. al [6]. The Boolean model allows to model and simulate a huge variety of textures e.g. for crystals, leaves, etc. The texture model X is obtained by taking various realizations of compact random sets, implanting them in Poisson points in R^n, and taking the supremum. The functional moment $Q(B)$ of X, after Booleanization, is calculated as:

$$P(B \subset X^c) = Q(B) = \exp(-\theta \overline{Mes}(X' \oplus \overset{\vee}{B})) \qquad \forall B \in \kappa \quad (9)$$

where κ is the set of the compact random set of R^n, θ the density of the process and $\overline{Mes}(X' \oplus \overset{\vee}{X})$ is an average measure that characterizes the geometric properties of the remaining set of objects after dilation. Relation (9) is the fundamental formula of the model. It completely characterizes the texture model. $Q(B)$ does not depend on the location of B, i.e., it is stationary. One can also provide that it is ergodic so that we can peak the measure for a specific portion of the space without referring to the particular portion of the space.

Formula 9 show us that the texture model depends on two parameters:

- the density θ of the process and
- a measure $\overline{Mes(X` \oplus \overset{\vee}{B})}$ that characterizes the objects. In the one-dimensional space it is the average length of the lines and in the two-dimensional space

$\overline{Mes(\check{X} \oplus \check{B})}$ is the average measure of the area and the perimeter of the objects under the assumption of convex shapes.

Table 3. Texture Features based on Random Set

Description	Name	Type	Formula
Area in class image *t*	***Area_t***	numerical	$Area_t = \begin{cases} Area_t = Area_t + 1 & \text{if } f(x,y,t) = 1 \\ Area_t = Area_t & \text{if } f(x,y,t) = 0 \end{cases}$
Density in class image *t*	***Dens_t***	numerical	$Dens_t = \begin{cases} Dens_t = Dens_t + \frac{1}{A} & \text{if } f(x,y,t) = 1 \\ Dens_t = Dens_t & \text{if } f(x,y,t) = 0 \end{cases}$ with $A = \sum_{t=1}^{S} Area_t$
Number of objects	***Count_t***	numerical	*n(t)*
Mean area of objects in class image *t*	***AreaMean_t***	numerical	$\overline{A(t)} = \frac{1}{n(t)} \sum_{i=1}^{n(t)} A_i(t)$
Standard deviation of the area of the objects in class image *t*	***AreaStdDev_t***	numerical	$S(t) = \sqrt{\frac{1}{n(t)} \sum_{i=1}^{n(t)} (A_i(t) - \overline{A(t)})^2}$
The contour length of a single object is $u = l + \sqrt{2} \cdot m$ with *l* being the number of contour pixels having odd chain coding numbers and *m* being the number of contour pixels having even chain coding numbers.			
Mean contour length of objects in class image *t*	***ContMean_t***	numerical	$\overline{u}(t) = \frac{1}{n(t)} \sum_{i=1}^{n(t)} u_i(t)$
Standard deviation of the contour length of objects in class image *t*	***ContStdDev_t***	numerical	$S(t) = \sqrt{\frac{1}{n(t)} \sum_{i=1}^{n(t)} (u_i(t) - \overline{u(t)})^2}$

We consider the two-dimensional case and develop a proper texture descriptor.

Suppose now that we have a texture image with 8bit gray levels. Then we can consider the texture image as the superposition of various Boolean models, each of them having a different gray level value on the scale from 0 to 255 for the objects within the bit plane.

To reduce the dimensionality of the resulting feature vector, the gray levels ranging from 0 to 255 are now quantized into S intervals t. Each image f(x,y) is classified according to the gray level into t classes, with t={0,1,2,..,S}. For each class a binary image is calculated containing the value “1” for pixels with a gray level value falling into the gray level interval of class t and value “0” for all other pixels. The resulting bit plane *f(x,y,t)* can now be considered as a realization of the Boolean model. The quantization of the gray level into *S* intervals was done at equal distancet. In the following, we call the image *f(x,y,t)* a class image. Object labeling is done in the class

images with the contour following method [7]. Afterwards, features from the bitplane and from these objects are calculated.

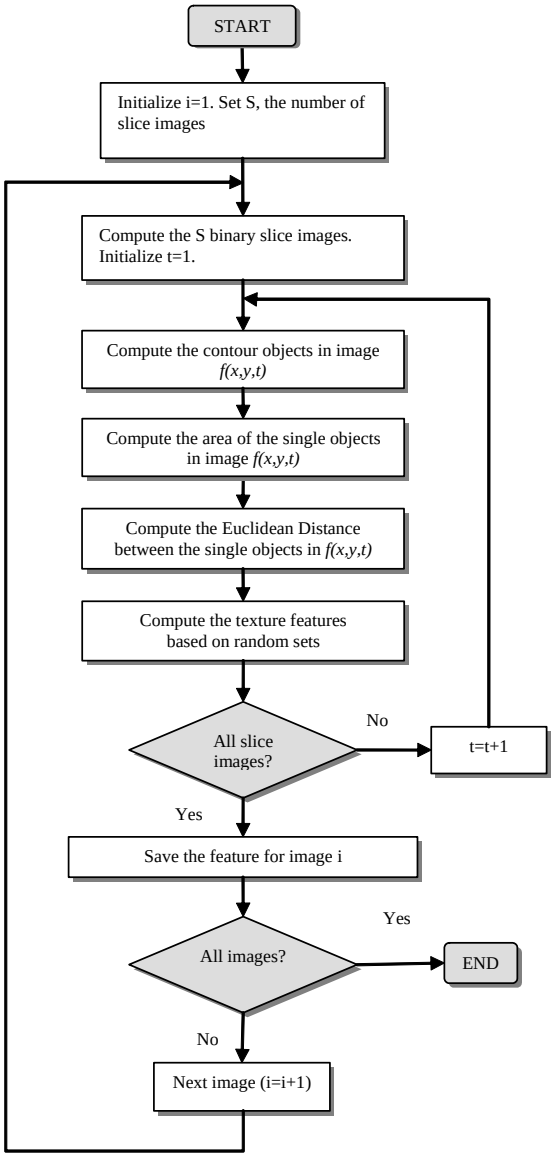

Fig. 1. Algorithm for the Texture descriptor based on Random Sets

The first one is the density of the class image *t* which is the number of pixels in the class image, labeled by “1”, divided by the area of the image. If all pixels of an image are labeled by “1”, then the density is one. If no pixel in an image is labeled, then the

density is zero. From the objects in the class image *t*, the area, a simple shape factor, and the length of the contour are calculated. According to the model, not a single feature of each object is taken for classification, but the mean and the variance of each feature are calculated over all the objects in the class image *t*. We also calculate the frequency of the object size in each class image *t*. The list of features and their calculation are shown in Table 3. The algorithm for the texture descriptor based on random sets is represented in Figure 1.

Depending on the number of slices S we get a feature set of 42(*S=6*), 84(S=12), 112(S=16).

3 Material and Application

We studied the performance of the two texture descriptors based on a data set of 344 images. These images come from an endoscopic video system used for colon examination [8]. The data set contains 283 normal tissue images and 61 polypimages (see Figure 2) in the form of sub-images of a size 33x33 that are derived from 37 original colonoscopic images. The polyps in the 37 original colonoscopic images were identified and selected by a "well-trained" medical expert. A polyp is split into as many as possible sub-images.

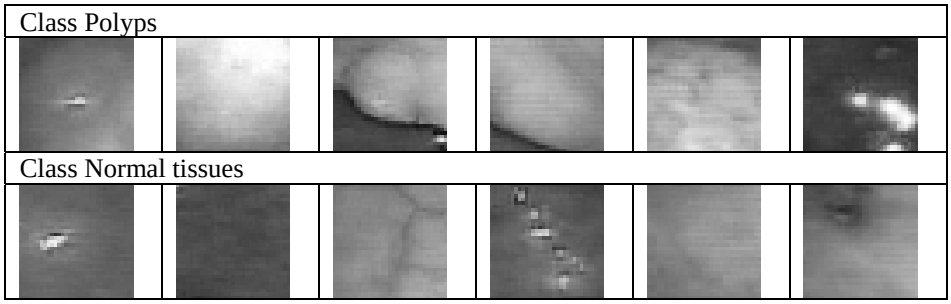

Fig. 2. Some Exemplary Images

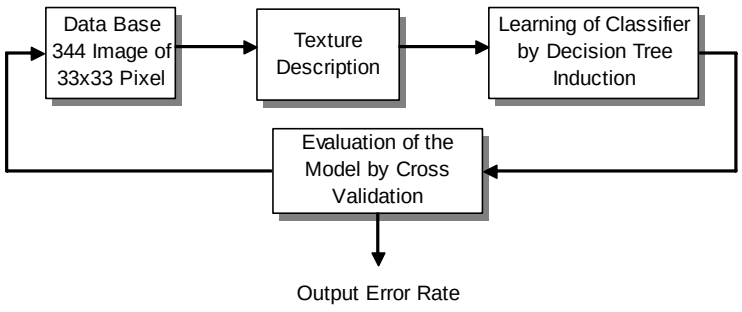

Fig. 3. Methodology

The 283 normal images consist of dark regions, reflections etc. of the 37 original colonoscopic images.

This presents a two class problem; one must decide if the image shows a polyp or not. The texture descriptions were calculated from these images. The resulting data set was used to train a decision tree based on the C4.5 algorithm [9]. Cross-validation was used to estimate the error rate.

4 Results

For the texture descriptor based on random sets the choice of S is important. On the one hand, we need a sufficiently large *S* to separate the classes. On the other hand, with increasing S also the number of features increases and we run into the curse-of-dimensionality problem.

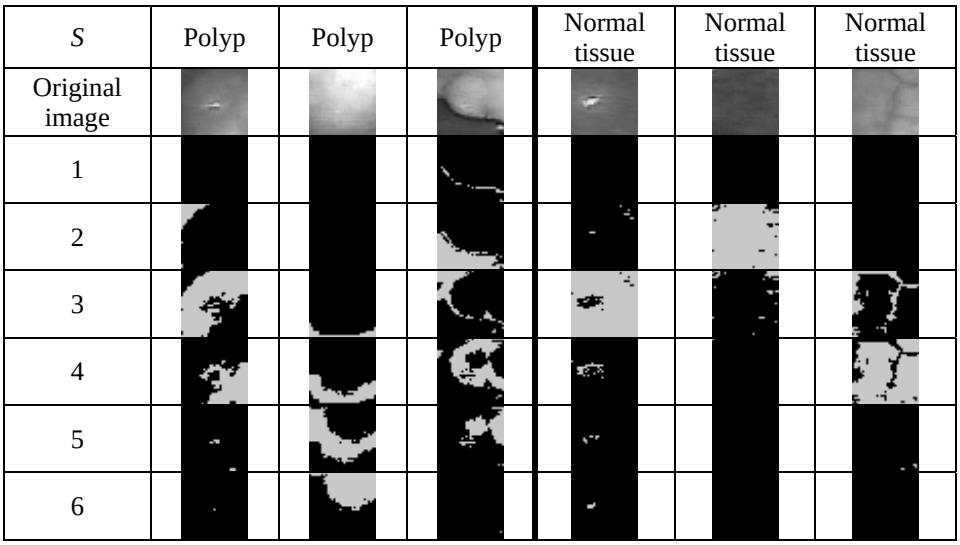

Fig. 4. The images f(x,y,t) with S=6

Figure 4 shows the class images for some polyp images and some normal tissue images for *S=6*. Figure 4 shows the class images for some polyp images and some tissue images for *S=12*. Figure 5 shows that most pixels of normal tissue images are located in only a few lower 1-3 class images. In contrast to this, in the polyp images the pixels are distributed more across the class images.

For our tests we used *S=6*, *S=12* and *S=16*. We have not yet developed a good procedure to estimate the number of *S*. The determination of the right number of *S* is still heuristic but in most of our applications S=12 turned out to be a good choice [4].

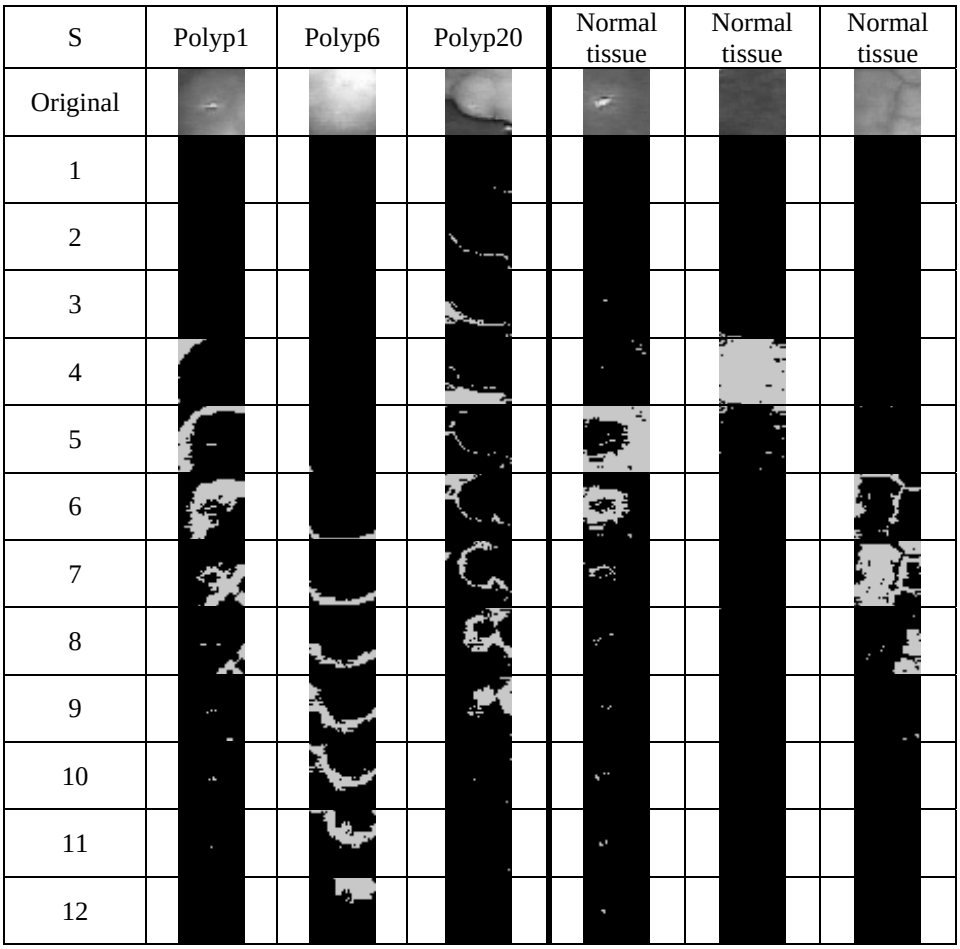

Fig. 5. The images f(x,y,t) with S=12

In the first test we used 30 polyp images and 30 normal tissue images as a data base. The results are shown in Figure 6.

In the second tests we used all 344 images as a data base. The results are shown in Figure 7.

In both tests the texture descriptor based on random sets with S=12 is the best texture descriptor. The test shows that the choice of *S=6* is too small and the choice of *S=16* is already too large. This observation might already demonstrate the effect of the curse of dimensionality.

The texture descriptor based on random sets for S=12 has an error rate of 1.67% for the data set with 60 images (see Figure 6) with equally distributed number of polyps and normal tissue. Compared to this, the texture descriptor Haralick-1 has an error rate of 3.33%, Haralick-2 has an error rate of 10%, and Haralick-3 has an error rate of 11.67%. Using all fifty two features as in Haralick-3 gives no improvement in

the quality of the decision model. It seems that there is a high correlation among these features and that the selected features as in Haralick-1 have turned out to be the most important ones in several studies.

The texture descriptor based on random sets for S=12 has an error rate of 9.88% for the data set with 334 images (see Figure 7) with 283 normal tissues and 61 polyps. Compared to this, the texture descriptor Haralick-1 has an error rate of 13.37% and Haralick-2 has an error rate of 18.89%.

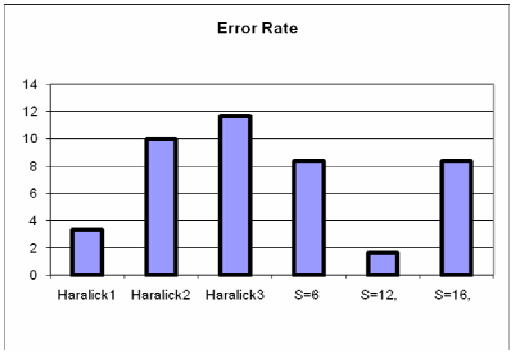

Fig. 6. Error rate (in percent) for Test 1

Fig. 7. Error rate (in percent) for Test 2

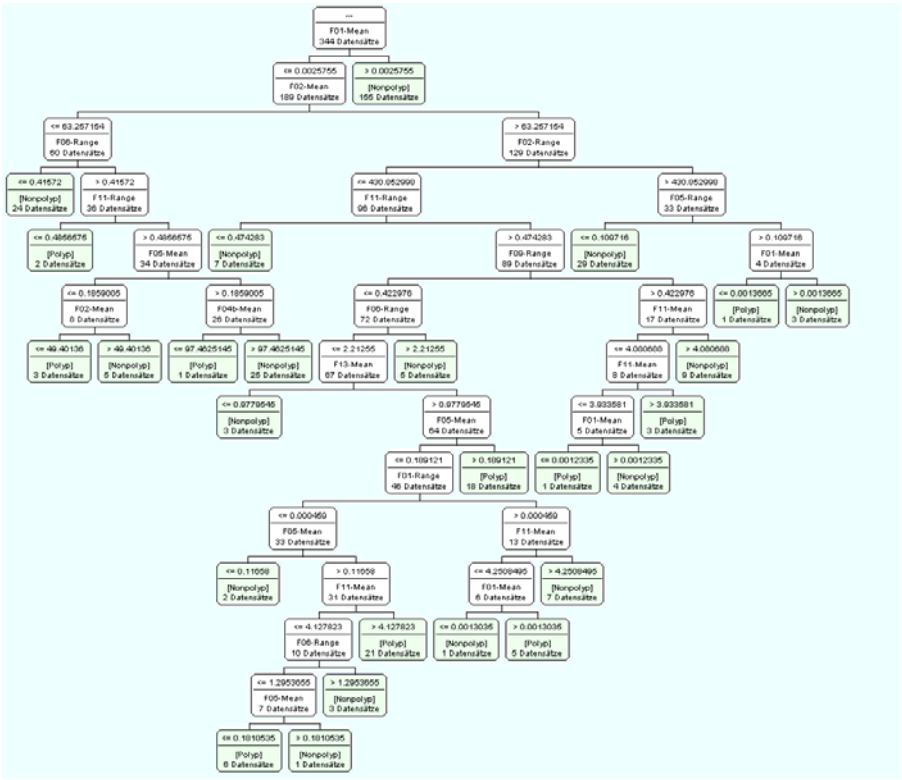

Fig. 8. Decision Tree for Haralick´s Feature Descriptor

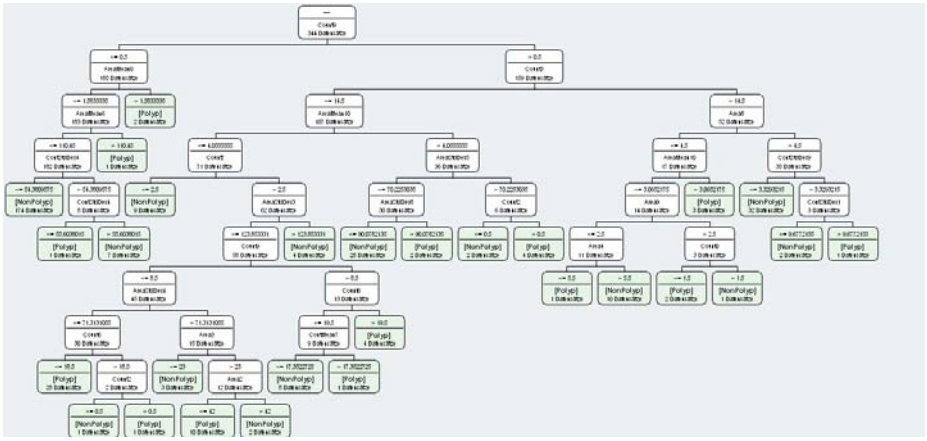

Fig. 9. Decision Tree for Texture Features based on Random Sets

Table 4. Selected Features, Decision Tree Properties, Run Time

	Haralick`Texture Descriptor		**Texture Descriptor based on Random Sets**
Selected Features	12		22
Names of Features	Mean: F01, F02, F04, F05, F11, F13 Range: F01, F02, F05, F06, F09, F11		AreaMean0, Count0 ContStdDev1 Area2, Count2 Area3, AreaStdDev3, Count3 Area4, AreaStdDev4, ContStdDev4 AreaStdDev5 Count6, AreaMean6 Count7, ContMean7 AreaStdDev8, Area8 Area9, Count9, ContStdDev9 AreaMean10
Width of Tree	8		8
Depth of Tree	13		10
Runtime	Haralick-1	Haralick-2	13.75s
	91.03s	83.22s	

The resulting decision trees are shown in Figure 8 for Haralick´s feature descriptor and in Figure 9 for the texture features based on random sets.

The comparison of the two trees shows that the feature selection method during decision tree induction selects only 12 features from 26 features for Haralick´s texture descriptor and 22 features from 84 features for the texture descriptor based on random sets (see Table 4). The tree expands more in depth for the Haralick´s feature descriptor than for the texture descriptor based on random sets. The runtime of the program for the calculation of Haralick´s texture descriptor is 7-times longer than for the texture descriptor based on random sets.

The runtime of the program for the calculation of Haralick´s-2 texture descriptor is not as long but the error rate is much higher than that for Haralick-1.

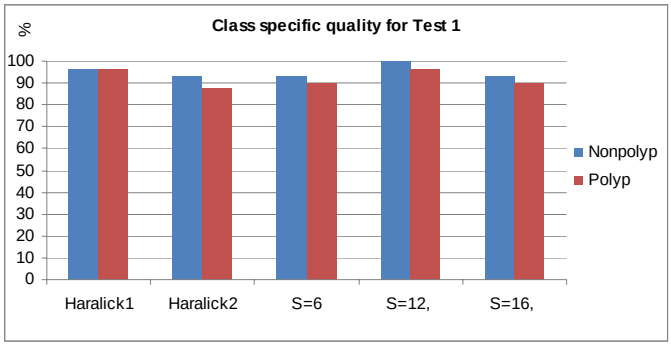

Fig. 10. Class specific quality for Test 1

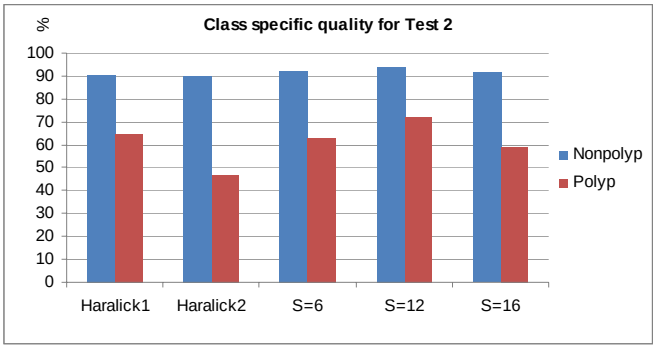

Fig. 11. Class specific quality for Test 2

The overall error rate is usually determined in many data mining experiments. Perner et. al. [10] have been shown that it is often necessary to determine more specific error rates in order to judge the quality of the classification model. Figure 10 and Figure 11 show the class specific quality. Both diagrams show that the texture descriptor based on random sets with S=12 gives the best results. In Figure 10 the two classes are equally good classified. In Figure 11 is the classification of nonpolyp images better than the one of polyp images. Physicans prefer to have a better class specific quality for the images that show the disease since it is better to have among the decisions one patient who is wrongly classified for having the disease and provide him some treatment than an unhealthy person gets classified as healthy. The results in Figure 10 are influence from the fact that we used in test2 an unbalanced data set (283 normal tissue images/ 61 polyp images). For example for the texture descriptor based on random sets with S=12 17 polyp images and 17 normal tissue image are misclassified in test. Nonetheless, the diagram shows clearly that also the class specific quality for the texture descriptor based on random sets is better.

5 Does Feature Preselection Improve the Model?

All our data sets have a large number of features. We expect to run into the curse of dimensionality. Therefore we used feature preselection before we run the decision tree experiment.

We use the feature selection method Relief [11] from the data mining tool Weka with the parameters:

- relief1: 10 Number of nearest neighbor: 10
- relief2: 10 Number of nearest neighbor: 3.

The results are shown in Figure 12. The diagram shows that except for S=12 we get better classification results for Random Sets for S=16. The improvement is up to 26.5 % for S=16 and relief1.

However, the results for S=12 remain stable and it seems that this data set is the right choice between the texture description and the quality of the classification model.

Fig. 12. Error rate (in percent) for Test 2

6 How to Estimate the Number of the Parameter S for the Texture Descriptor Based on Random Sets?

The estimation of the parameter S for the Texture Descriptor based on Random Sets is still done empirically. We like to find a method that allows us to calculate this parameter in advance.

The main idea of the Random Set Texture Descriptor is to quantize the gray level histogram into bins. We have a domain that is a function $f_{256}(x) : x \in \mathbb{N}$ and $x \in [1, 256]$

x	1	2	…	255	256
$f_{256}(x)$	f_1	f_2	…	f_{255}	f_{256}

We transform the function to the following function $g_s(y)$ and $i \in \mathbb{N}$:

y	y_1	y_2	…	y_{S-1}	y_S
$g_S(y)$	$g_1 = \sum_{i=1}^{i<1+256/S} f_i$	$g_2 = \sum_{i=\lceil 1+256/S \rceil}^{i<1+2*256/S} f_i$	…	$g_{S-1} = \sum_{i=\lceil 1+(s-2)*256/S \rceil}^{i<1+(S-1)*256/S} f_i$	$g_S = \sum_{i=\lceil 1+(S-1)256/S \rceil}^{i<257} f_i$

This step can be seen as approximation of the histogram for which we need to know what is the right number of bins [12]. Several methods for computing the optimal number of bins for a histogram from a set of data are known from the literature [12-15].

Let S be the number of the bins, h the width of the bins and n the number of data. In our case the data are the number of pixels.

The following rules can be used to estimate the number of bins:

- Sturges formula [13]:

$$k = \lceil 1 + \log_2 n \rceil \tag{10}$$

- Square-root choice [14]:

$$k = \sqrt{n} \tag{11}$$

- Dreyer and Sauer [15]:

$$k = 5 \lg n \,. \tag{12}$$

Each of the rules is a compromise between information reduction and clarity [14]. Furthermore the methods make different assumptions for the shape of the distribution. For example the rule from Sturges [13] is only for normal distributed data set.

Table 5 shows the calculated number of bin using the formulas 10-12. These formulas take only the number of data (in our cases number of pixels) into account.

Table 5. Different Estimates for number of bins

Name of Rule	Polyp Images
Sturges	12
Dreyer/ Sauer (p. 149)	15
Square-root choice	33

It shows that the Sturges rule comes close to our result of S=12 while the rule given by Dreyer and Sauer estimates a much larger S=15, which might also be applicable but needs feature preselection to obtain a good classification model. The square-root rule estimates a too large interval and is not applicable to our application.

The results show that transformation of the image into class images has something to do with histogram approximation. The question is: Does the texture descriptor require a low approximation failure of the gray level histogram or should the quantization of the gray level be done in such a way that each class image contains sufficient information content. This question will be left for further work. The images and the histograms in Figures 13 and 14 give a hint to think into the direction that the the texture descriptor might require the quantization of the gray level be done in such a way that each class image contains sufficient information content.

7 Discussion

In this application the texture descriptor based on random sets outperformed Haralick´s texture descriptor. The accuracy is 3.49 % higher than that of Haralick´s texture descriptor in case of Haralick-1 and 9.01% higher in case of Haralick-2.

Decision trees are sensitive to unbalanced class distribution. Therefore, the error rate in the second experiment rises since the ratio of the two classes is 1/5 in the data set. Nonetheless, the tendency of the error rate of the three descriptors is the same.

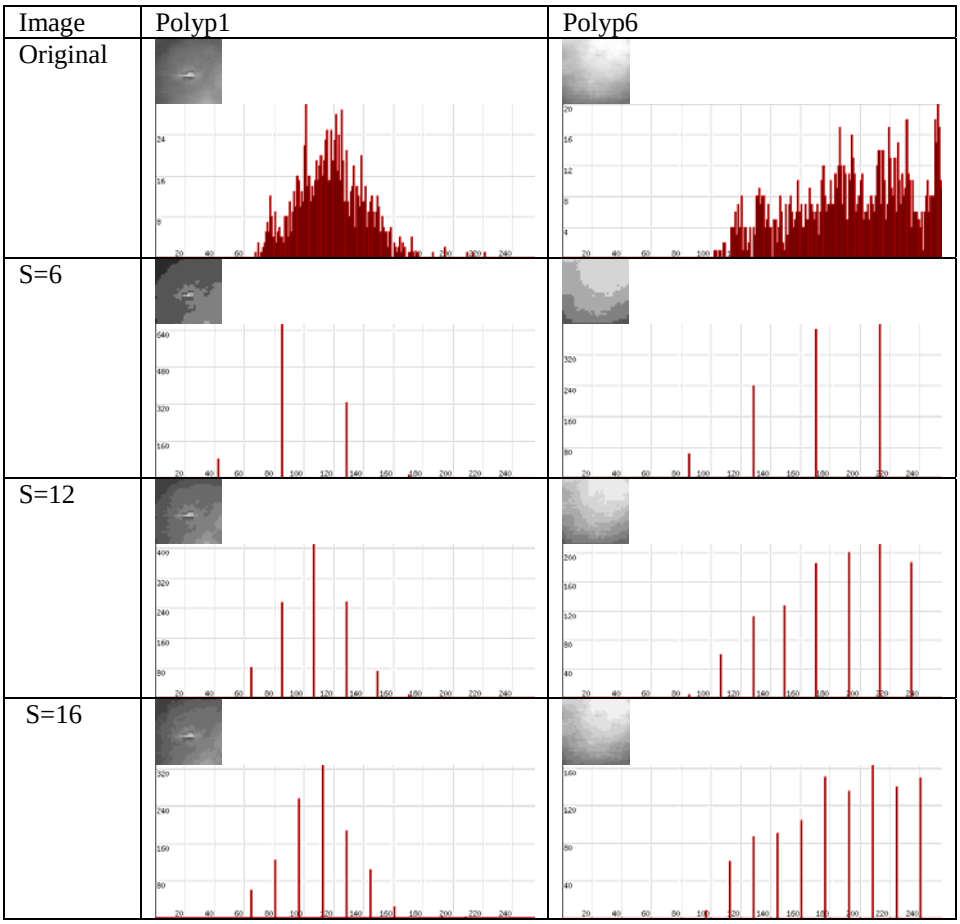

Fig. 13. Images and Histograms for some polyp images

A further advantage of the texture descriptor based on random sets over Haralick's texture descriptor is the reduced time required for computing the features. In addition, we can understand the semantics behind the numerical texture description. The texture features based on random sets have a semantic meaning and give an expert an understanding about texture.

Both texture features result in a large number of features. Feature preselection before decision tree induction can improve the quality of the model but increases the cost for computation. Therefore it is more preferable to select a feature descriptor that does not require this step. The feature descriptor based on Random Sets and S=12 is one such decsriptor.

The quality of the model should not only be evaluated based on the overall error rate. More specific quality measures such as the class specific quality should also be estimated and give a more specific insight into the quality of the model.

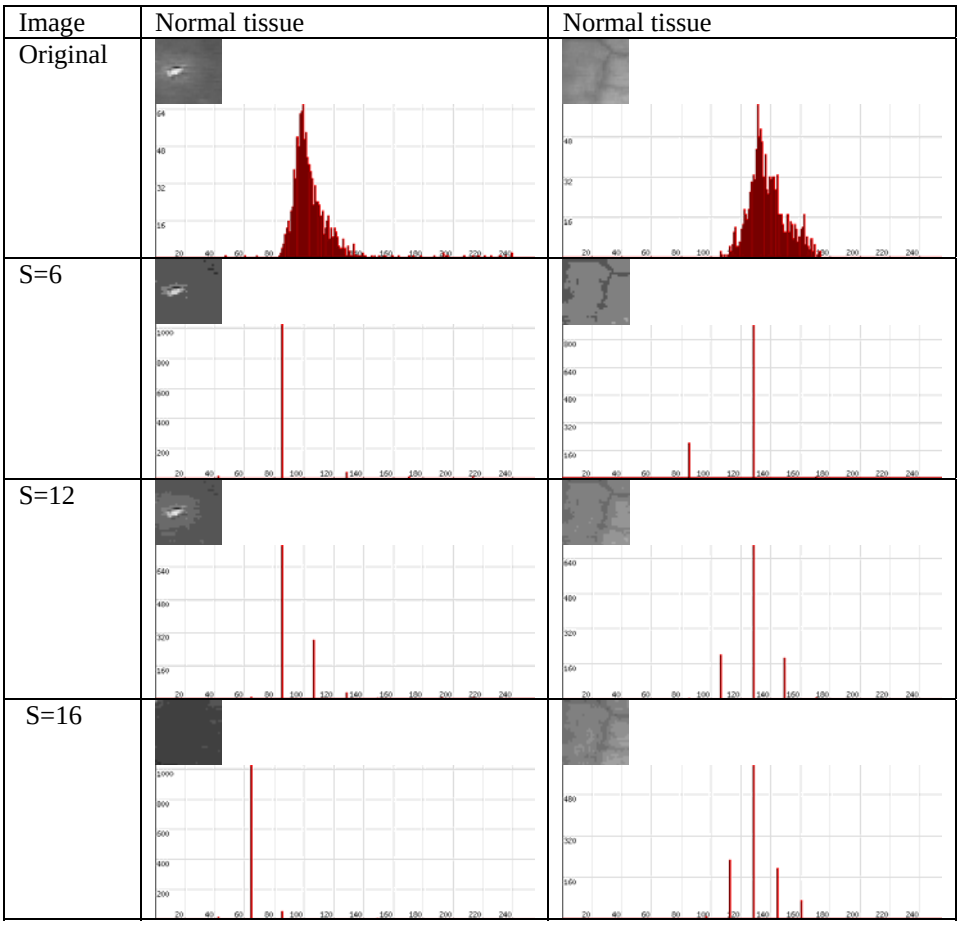

Fig. 14. Images and Histograms for some normal tissue images

The choice of the number of slices S is still empirical. In this paper we present a first idea about a procedure that allows us to estimate S. Based on our empirically tests and this rule we have found that S=12 gives the best results. The number S=12 provides a feature set of 84 features. It might be that this is a compromise between a rich description of texture and the large feature set problem (curse of dimensionality). We will continue to investigate this further.

The decision tree induction method performs feature selection during the tree building process. Therefore, the method can also be seen as a feature selector. The number of features selected for Haralick`s texture descriptor is always lower than the number selected for the texture descriptor based on random sets. The texture descriptor based on random sets may provide a more richer description of texture.

8 Conclusion

Medical disease examination is often based on images. Mining these images in order to obtain the classification knowledge for automatic image classification is a challenging task. This task belongs to the field of image mining. Image mining is usually not only comprised of mining a table of numbers it has also to do with transforming the image in the right image description. Both, the image description and the classification knowledge, determine the quality of the classifier.

Texture is a powerful method to describe biological objects. Many texture descriptors are known from the literature [1]. The most used texture descriptor is Haralick`s texture descriptor based on the co-occurrence matrix. We proposed a texture descriptor based on random sets [4] and in this paper compare both texture descriptors based on images that we derived from colon examination. We learnt a classifier model based on decision trees. Then we compared both texture descriptors based on the error rate, the class specific error rate, the number of selected features by the decision tree induction process, the tree properties of the learnt tree, and the runtime for the calculation of the texture descriptors.

Although decision tree induction can be seen as a feature selector in case of large number of features this procedure has some limitations. We show that feature preselection before decision tree learning can improve the model.

We have found that the texture descriptor based on random sets outperforms Haralick´s texture descriptor based on the error rate, tree properties and the runtime. Harlick´s texture descriptor uses fewer features from the set of calculated texture features than the texture descriptor based on random sets. However, this might only demonstrate that Haralick´s texture descriptor has limited description power since the error rate is much higher than that for the texture descriptor based on random sets.

In addition, the texture descriptor based on random sets has semantic meanings. An expert can understand the properties of a texture when looking into the slice produced during the calculation of the texture features. A problem still to be solved is how to predict the optimal number of slices. We provide a first rule for estimating the number of slices and compare the results with our empirically determined results. Further work will focus on this problem.

References

1. Rao A. R.: A Taxonomy for Texture Description and Identification, Springer Verlag, Berlin (1990)
2. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural Features for Image Classification. IEEE Transactions on Systems, Man and Cybernetics 3(6), 610-621 (1973)
3. Perner, P., Perner, H., Müller, B.: Texture Classification based on Random Sets and its Application to Hep-2 Cells. , In: Kasturi, R., Laurendeau, D., Suen, C. (Eds.) ICPR 2002, Vol. II, pp. 406-411, IEEE Computer Society (2002)
4. Perner P., Perner H., Müller B.: Mining Knowledge for Hep-2 Cell Image Classification, Journal Artificial Intelligence in Medicine (26), 161-173 (2002)

5. Matheron, G.: Random Sets and Integral Geometry. J. Wiley&Sons, New York, London (1975)
6. Stoyan, D., Kendall, W.S., Mecke, J.: Stochastic Geometry and Its Applications. Akademie Verlag (1987)
7. Zamperoni, P.: Methoden der digitalen Bildverarbeitung, 2. Auflage. Vieweg Verlag, Braunschweig (1991)
8. Cheng, D.C., Ting, W.C., Chen, Y.F., Pu, Q., Jiang, X.: Colorectal Polyps Detection Using Texture Features and Support Vector Machine. In: Perner, P., Salvetti, O. (eds.) MDA 2008. LNCS, vol. 5108, pp. 62-72, Springer (2008)
9. Data Mining Tool *Decision Master*, ibai-solutions www.ibai-solutions.de
10. Perner, P., Zscherpel, U., Jacobsen, C.: A Comparison between Neural Networks and Decision Trees based on Data from Industrial Radiographic Testing. Pattern Recognition Letters 22, 47-54 (2001)
11. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P.,. Witten, I.H.: The WEKA Data Mining Software: An Update; SIGKDD Explorations 11 (1), 10-18 (2009)
12. Freedman, D., Diaconis, P.: On the histogram as a density estimator: L_2 theory. Probability Theory and Related Fields 57 (4), 453-476 (1981)
13. Sturges, H.A.: The choice of a class interval. Jornal of the American Statistical Association 21, 65-66 (1926)
14. Schaich, E.: Schätz- und Testmethoden für Sozialwissenschaftler, 2. Auflage, Verlag Vahlen, München (1990)
15. Dreyer, H., Sauer, W.: Prozeßanalyse. VEB Verlag Technik, Berlin (1982)