Big Data, Decision Tree Induction, and Image Analysis for the Discovery of Decision Rules for Colon Examination

Petra Perner

Institute of Computer Vision and Applied Computer Sciences, IBaI, Leipzig, Germany

Abstract— *The aim of our research was to develop a method that allows us automatically to discover the decision rules for diagnosing medical images in normal tissue images and images showing a polyp. We used a data set of images that came from an endoscope video system used for colon examination. The data set contains 283 normal tissue images and 61 polyp images. The 283 normal images consist of dark regions and reflection. One must decide if the image shows a polyp or not. This is a two-class problem. The unequal number of the data in the two classes makes our problem to an unbalanced data set problem. The polyps in the images were identified and selected by a "well-trained" medical expert. Based on these medical images, we study the behavior of two different statistical texture descriptors, the co-occurrence matrix-texture descriptor and our novel Random set texture descriptor. We review the theory of both texture descriptors and then we apply them to our medical data set. We used a decision-tree induction method to learn the classification rules based on our tool "Decision Master". In both cases, for the full unequally distributed data set and for the balanced data set, we achieved the best error rate based the Random-set texture descriptor. The performance of the co-occurrence matrix-texture descriptor was worse. For statistical based texture descriptors large enough texture are necessary that cannot always guaranteed for medical objects. Since the co-occurrence matrix is based on higher order statistic that might be the reason for the worse performance. The results show that decision tree induction and image analysis based on our novel texture descriptor is an excellent method to mine medical images for the decision rules even when the data set is unbalanced, but not only that makes our Random-set based texture descriptor favorable. It also gives a flexible way to describe the appearance of the medical objects in symbolic terms, the computation time is less, and it can be set up as software module that can be flexible used in different systems.*

Keywords— Image Analysis, Endoscope Images, Colon Examination, Polyp Images, Decision Tree Induction, Random Set Texture Descriptor, Co-occurrence Texture Descriptor, Unbalanced Data Set Problem.

I. INTRODUCTION

The aim of our research was to develop a method that allows us automatically to discover the decision rules for diagnosing images in normal tissue images and images containing a polyp. We used a data set of medical images that came from an endoscope video system used for colon examination. Texture seems to be a powerful tool to describe the appearances of medical objects into normal tissue and polyp´s. Therefore, very flexible and powerful texture descriptors are of importance that allows to recognize the texture and to understand what makes up the texture. Texture seems to become an important role to describe the appearance of different medical and 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 (Rao 1990). The most used texture descriptor is the well-known texture descriptor based on the co-occurrence matrix (Haralick et. al 1973). Although it works well on different applications we prefer to use our texture descriptor based on Random sets (Perner et. al 2002) since this descriptor gives us more freedom in describing different textures. In this paper, we compare the two texture descriptors based on a medical data set. Related work on texture description is given Section 2. The theory of the texture descriptors based on the co-occurrence matrix is reviewed in Section 3 and the texture descriptor based on Random sets is reviewed in Section 4. The material and the application of the texture descriptors and the decision-tree induction-method is explained in Section 5. The used data set of polyp 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 6 and they are discussed in Section 7. Conclusions are given in Section 8.

II. RELATED WORK

Texture description methods are mainly classified into structural, statistical, model-based, and transform-based approaches

(Bharatiet. al 2004, Castellano et. al 2004, Zhang & Tan 2002). Structural methods use texture elements to describe textures. It is good for image synthesis applications. Statistical methods use gray-level relationship between neighboring pixels to describe to local texture property in first-order, second-order, or higher-order statistics. The methods are good for invariant texture analysis and classification. Model-based methods model images as different probability or linear combination models (Zhang & Tang 2002) and use model parameters to describe their texture features, such as autoregressive models, fractal models (Kaplan 1999), Gaussian-mixture models (GMM) (Nguyen & Wu 2012), hidden Markov models (HMM) (Chen & Kundu 1993, Najmi & Gray 2000), Markov random fields (MRF) (Krishnamachari & Chellappa 1997) and so on. The transform methods transfer images into a frequency domain to describe textures. The methods usually use Fourier, Gabor, or wavelet transform. An overview about older methods such autocorrelation and other is given in Haralick (1979) and van Gool et. al (1985).

Often the texture descriptors are compared on standard texture data sets but recently appeared work of texture description for real world problems such as description of objects in medical images, microscopic images for different purposes such as e.g. in system biology and for environmental applications, food inspection and so on. Texture became valuable information about images. Researcher tries to develop many new texture descriptors that take into account the variances of the texture, the spectral influences and so on. At lot of different methods exist and it is not easy to do a categorization of all these methods. We want to describe in brief the recent developments. They are often variants of the above-described categories that have been evaluated on standard data sets. However nowadays, more work on real world applications appear.

Din-Chang Tseng et. al(2015) developed a multiscale texture segmentation approach based on contextual hidden Markov tree (CHMT) model and boundary refinement. A hidden Markov tree (HMT) model is a probabilistic model for capturing persistence properties of wavelet coefficients without considering clustering properties. They have proposed the CHMT model to enhance the clustering properties by adding extended coefficients associated with wavelet coefficients in every scale.

Wesley Nunes Goncalves et. al(2014) developed a method that can capture the details richness of the image surface. They estimated the fractal dimension by the Bouligand- Minkowski method due to its precision in quantifying structural properties of images. They validated their method on two standard texture datasets and the experimental results reveal that the methods are good enough to describe different data sets.

R. Mukundan (2014) use orthogonal moment functions based on Tchebycheff polynomials. They claim that the method is good because of their superior feature representation capabilities. They construct feature vectors from orthonormal Tchebycheff moments evaluated on 5x5 neighborhoods of pixels, and encoding the texture information as a Lehmer code that represents the relative strengths of the evaluated moments. The features will be referred to as Local Tchebycheff Moments (LTMs). The encoding scheme provides a byte value for each pixel, and generates a gray-level _LTM-image_ of the input image. The histogram of the LTM-image is then used as the texture descriptor for classification.

YuhuiQuan et. al(2014) developed a statistical approach to static texture description, which combines a local pattern coding strategy with a robust global descriptor to achieve highly discriminative power, invariance to photometric transformation and strong robustness against geometric changes. They called their method pattern fractal spectrum that characterizes the self-similar behavior of the local pattern distributions by calculating fractal dimension on each type of pattern. Compared with other fractal-based approaches, the proposed descriptor is compact, highly distinctive and computationally efficient. The evaluation was done on a standard benchmark set.

Aujol et. al(2006) explored in their paper various aspects of the image decomposition problem using modern variational techniques. They aim at splitting an original image f into two components u and v, where u holds the geometrical information and v holds the textural information. The modeling uses the total-variation energy for extracting the structural part and one of four of the following norms for the textural part: L2, G, L1 and a new tunable norm, suggested there for the first time, based on Gabor functions. They design tools for the TV -Gabor model.

Champion et. al(2014) do texture modelling on a real-world application for forest stand age from SAR images. The texture descriptors are calculated from statistics generated by the gray-level co-occurrence matrix for varying distance d, and orientation α, values used to calculate the matrix. It is found that texture descriptors such as contrast, inverse-difference moment, homogeneity, and correlation are strongly influenced by the parameters (d, α) related to forest stand structure (forest rows, stand density) and image resolution. In contrast, the calculated energy and entropy from the co-occurrence matrix are observed to be highly correlated to stand age and displayed a stable performance whatever the distance and orientation

parameters (d, α), thus rendering them a good contender.

Dharmagunawardhana et. al(2014) proposed a novel robust texture descriptor based on Gaussian Markov random fields (GMRFs). A spatially localized parameter estimation technique using local linear regression is performed and the distributions of local parameter estimates are constructed to formulate the texture features. The inconsistencies arising in localized parameter estimation are addressed by applying generalized inverse, regularization, and an estimation window-size selection criterion. The texture descriptors are named as local parameter histograms (LPHs) and are used in texture segmentation with the k-means clustering algorithm. The segmentation results on general texture datasets demonstrate that LPH descriptors significantly improve the performance of classical GMRF features and achieve better results compared to the state-of-the-art texture descriptors based on local feature distributions.

Madzin et. al(2014) deal with medical application, where the usage of multiple medical images generated by computer tomography such as x-ray, Magnetic Resonance Imaging (MRI) and CT-scan images is a standard tool of medical procedure for physicians. The major problems in analyzing various modality of medical image are the inconsistent orientation and position of the body-parts of interest. In this research, local descriptor of texture, shape and color are used to extract features from multi-modality medical image in patches and interest point's descriptor.

Palanivel et. al (2015) use a Markov process with Bayesian Approach to analyze textures in the image and that are identified and distinguished from untextured regions with edges. The parameters of the model are estimated based on the Bayesian approach. They use two types of classification namely supervised and unsupervised classification.

Massich et. al (2014) use Self-Invariant Feature Transform (SIFT), both as low-level and high-level descriptors, applied to differentiate the tissues present in breast US images. For the low-level texture descriptors case, SIFT descriptors are extracted from a regular grid. The high-level texture descriptor is built as a Bag-of-Features (BoF) of SIFT descriptors. Experimental results are provided showing the validity of the proposed approach for describing the tissues in breast US images.

Song et. al (2014) presented a noise-robust descriptor by exploring a set of local contrast patterns (LCPs) via global measures for texture classification. To handle image noise, the directed and undirected difference masks are designed to calculate three types of local intensity contrasts: directed, undirected, and maximum difference responses. To describe pixel-wise features, these responses are separately quantized and encoded into specific patterns based on different global measures. These resulting patterns (i.e., LCPs) are jointly encoded to form our final texture representation. The evaluation has been done on two standard data sets and showed superior performance compared too many state-of-the-art methods.

Zhang and Pham (2010) and Pham (2014) tried to recognize the Subcellular Location Features (SLF) by three well-known texture feature descriptions, which are the local binary patterns (LBP), Gabor filtering and Gray Level Co-occurrence Matrix (GLCM), to recognize the cell phenotype images. Using the public benchmark 2D HeLa cell images, high classification accuracy 96% is obtained with rejection rate 21% from the proposed system by taking advantages of the complementary strengths of feature construction and majority-voting based classifiers' decision fusions.

Marcos et. al (2015) use Gray-Level Co-occurrence Matrices (GLCM), Log-Gabor Filters (LGF), Local Binary Patterns (LBP) and Discrete Tchebycheff Moments (DTM) for pollen identification in microscopic images. Fisher's discriminant analysis and k-nearest neighbor were subsequently applied to perform dimensionality reduction and multivariate classification, respectively. They found that the combination of all the texture features resulted in the highest performance, yielding an accuracy of 94.83%.

Olveres et. al (2014) use texture image segmentation for medical images. The noise inherent to images and the lack of contrast information between adjacent regions hamper the performance of the algorithms. The characterizations of regions as statistical parametric models to handle level set evolution have been proposed. In this paper, they study the influence of texture on a level-set-based segmentation and propose the use of Hermite features that are incorporated into the level set model to improve organ segmentation that may be useful for quantifying left ventricular blood vessel. The proposal was also compared against other texture descriptors such as local binary patterns, Image derivatives, and Hounsfield low attenuation values.

Cai et. al(2015) propose a novel phase-based texture descriptor for efficient and robust classifiers to discriminate benign and malignant tumors in breast cancer images. The phased congruency-based binary pattern (PCBP) is an oriented local texture descriptor that combines the phase congruency (PC) approach with the local binary pattern (LBP). The proposed PCBP texture descriptor achieves the highest values (i.e. 0.894) and the least variations in respect of the AUC index, regardless of

the gray-scale variations.

Cheng et. al (2008) propose a texture method based on the co-occurrence matrix to detect colorectal polyps in colonoscopy images. They used support vector machines for classification and achieve a sensitivity of 86, 2%.

We have developed our own texture descriptor based on statistics that model the texture by a Poisson process after the image is processed by a morphological operation. The remaining areas in the images can be described by first-order and second-order statistics as well as higher-order statistics if the numbers of remaining areas are large enough. The texture descriptor can be easily and fast computed and can handle different medical textures very well (Perner 1999, Perneret. al 2002). These medical textures are often not easy to describe as it is in case of the Brodatz texture data set[1]. Our method has also explanation capability. A human can understand the differences in the texture by looking up the remaining images. If necessary, a symbolic description of the different textures can be found. Our texture descriptor has still some other properties that are of interest but here in this paper, we want to compare our texture descriptor to the co-occurrence matrix since it is from the category of statistic texture descriptors. The co-occurrence matrix is still the most used texture descriptor and we want to explore the differences between our texture descriptors and the co-occurrence matrix.

III. THE CO-OCCURRENCE TEXTURE DESCRIPTOR

The co-occurrence texture feature descriptor is comprised of fourteen statistical features (Haralick et. al 1973) derived from the co-occurrence matrix.

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*:

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

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.

We can define the normalized co-occurrence matrix from $C_{(\Delta x,\Delta y)}$ as:

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

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 for the line respectively.

$$p_x(i) = \sum_{j=1}^{N_g} p_{ij} \quad \text{and} \quad p_y(i) = \sum_{i=1}^{N_g} p_{ij} \tag{3}$$

Further, we are calculating:

$$p_{x+y}(k) = \sum_{\substack{i=1 \\ i+j=k}}^{N_g}\sum_{j=1}^{N_g} p_{ij} \tag{4}$$

With *k*=2, 3,. . . .2 N_g

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

This results in:

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

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

Now we can define our co-occurrence texture features by well-known discreet chance dimensions such as the average, moments, and variants, measures that describe the dependence of two random variables such as the correlation (Dreyer & Sauer 1982), and the measure of the mess such as the entropy:

1. Angular Second Moment:

$$f_1 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_{ij}^2 \tag{8}$$

2. Contrast

$$f_2 = \sum_{k=0}^{N_g} k^2 \left(\sum_{\substack{i=1 \\ |i-j|=k}}^{N_g}\sum_{j=1}^{N_g} p_{ij} \right) \tag{9}$$

3. Correlation

$$f_3 = \frac{1}{\sigma_x \sigma_y}\left(\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} ijp_{ij} - \mu_x\mu_y \right) = \frac{1}{\sigma^2}\left(\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} ijp_{ij} - \mu^2 \right) \tag{10}$$

4. Sum of Squares Variance

$$f_4 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} (i-\mu)^2 p_{ij} \tag{11}$$

5. Inverse Difference Moment

$$f_5 = \sum_{i=1}^{N_g}\sum_{j=1}^{N_g} \frac{1}{1+(i-j)^2} p_{ij} \tag{12}$$

6. Sum Average

$$f_6 = \sum_{k=2}^{2N_g} kp_{x+y}(k) \tag{13}$$

7. Sum Variance

$$f_7 = \sum_{k=2}^{N_g} (i-f_6)^2 p_{x+y}(k) \tag{14}$$

8. Sum Entropy

$$f_8 = -\sum_{k=2}^{2N_g} p_{x+y}(k) \log p_{x+y}(k) \tag{15}$$

9. Entropy

$$f_9 = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_{ij} \log p_{ij} \tag{16}$$

10. Difference Variance

$$f_{10} = \text{variance of } p_{x-y} \tag{17}$$

11. Difference Entropy

$$f_{11} = -\sum_{k=0}^{N_g-1} p_{x-y}(k) \log p_{x-y}(k) \tag{18}$$

12.

and

13. Information Measure of Correlation

$$f_{12} = \frac{f_9 - HXY1}{H} \tag{19}$$

$$f_{13} = \sqrt{1-\exp[-2(HXY2 - f_9)]\}} \tag{20}$$

Where

$$HXY1 = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_{ij} \log\left(p_x(i)p_y(j)\right) \tag{21}$$

$$HXY2 = -\sum_{i=1}^{N_g}\sum_{j=1}^{N_g} p_x(i)p_y(j) \log\left(p_x(i)p_y(j)\right) \tag{22}$$

$$H = \sum_{k=1}^{N_g} p(k) \log p(k) \tag{23}$$

For our tests, we compute four co-occurrence matrixes $C_{(\Delta x,\Delta y)}$ with the offsets shown in Table 1.

TABLE 1

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 thirteen co-occurrence texture features of formula (8)-(20). Thus, we obtain four values for each feature. To reduce the feature set, we compute analog to Haralicket. al (1973) the mean and rank of each feature so that we finally get twenty-six features. This feature set is named COO-1.

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

$$CT_{ij} = C_{(0,1)}(i,j) + C_{(-1,1)}(i,j) + C_{(-1,0)}(i,j) + C_{(-1,-1)}(i,j) \tag{24}$$

The matrix CT is normalized per formula 2. From the normalized matrix PT, we calculate the thirteen texture features of co-occurrence descriptor. This feature set is named COO-2.

In addition to the above described features, is defined the maximal correlation coefficient as feature number fourteen. However, we are not using this coefficient in our study because of the high computation time.

IV. TEXTURE DESCRIPTOR BASED ON RANDOM SETS

Boolean sets were invented by Matheron (1975). An in-depth description of the theory can be found in Stoyan et al(1987). 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:

$$P(B \subset X^c) = Q(B) = \exp(-\theta \overline{Mes}(X' \oplus \check{B})) \qquad \forall B \in \kappa \tag{25}$$

where κ is the set of the compact random set of R^n, θ the density of the process and $\overline{Mes}(X' \oplus \check{X})$ is an average measure that characterizes the geometric properties of the remaining set of objects after dilation. Relation (25) 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 25 show us that the texture model depends on two parameters:

- the density θ of the process and
- a measure $\overline{Mes(X` \oplus \check{B})}$ that characterizes the objects. In the one-dimensional space, it is the average length of the lines and in the two-dimensional space $\overline{Mes(X` \oplus \check{B})}$ are the average measure of the area and the perimeter of the objects under the assumption of convex shapes.

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

Suppose now that we have a texture image with 8 bit 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 distances. In the following, we call the image *f(x,y,t)* a class image. In the class image we can see a lot of different objects. These objects get labeled with the contour following method (Klette & Zamperoni 1996). Afterwards, features from the bit-plane and from these objects are calculated. Since it does not make sense to consider the features of every single object due to the curse of dimensionality, we calculate the mean and standard deviation for each feature that characterizes the objects such as the area and the contour. In addition to that, we calculate the number of objects and the areal density in the class image.

The list of features and their calculation are shown in Table 2. The first one is the areal 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.

TABLE 2
TEXTURE FEATURES BASED ON RANDOM SET

Description	Name	Type	Formula
Area in class image *t*	***Area_t***	num	$Area_t = \begin{cases} Area_t = Area_t + 1 & \text{if } f(x,y,t) = 1 \\ Area_t = Area_t & \text{if } f(x,y,t) = 0 \end{cases}$
Density in class image *t*	***Dens_t***	num	$Dens_t = \begin{cases} Dens_t = Dens_t + \frac{1}{A} & \text{if } f(x,y,t) = 1 \\ Dens_t = Dens_t & \text{if } f(x,y,t) = 0 \end{cases}$ with $A = \sum_{t=1}^{S} Area_t$
Number of objects	***Count_t***	num	*n(t)*
Mean area of objects in class image *t*	***AreaMean_t***	num	$\overline{A(t)} = \frac{1}{n(t)} \sum_{i=1}^{n(t)} A_i(t)$
Standard deviation of the area of the objects in class image *t*	***AreaStdDev_t***	num	$S(t) = \sqrt{\frac{1}{n(t)} \sum_{i=1}^{n(t)} (A_i(t) - \overline{A(t)})^2}$
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***	num	$\overline{u}(t) = \frac{1}{n(t)} \sum_{i=1}^{n(t)} u_i(t)$
Standard deviation of the contour length of objects in class image *t*	***ContStdDev_t***	num	$S(t) = \sqrt{\frac{1}{n(t)} \sum_{i=1}^{n(t)} (u_i(t) - \overline{u(t)})^2}$

From the objects in the class image *t*, the area, a simple shape factor, and the length of the contour are calculated. Per the model, not a single feature of each object is taken for classification due to the curse of dimensionality, but the mean and the standard deviation 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.

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

V. MATERIAL OF THE APPLICATION AND THE DATA-MINING TOOL *DECISION MASTER*®

We studied the performance of the two texture descriptors based on a data set of 344 images. These images come from an endoscope video system used for colon examination (Cheng et. al 2008). The data set contains 283 normal tissue images and 61 polyp images (see Figure 1) in the form of sub-images of a size 33x33 that are derived from 37 original colonoscopy images. The polyps in the 37 original colonoscopy images were identified and selected by a "well-trained" medical expert. A polyp is split into as many as possible sub-images.

The 283 normal images consist of dark regions, reflections etc. of the 37 original colonoscopy 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 (Perner 2002). Cross-validation was used to estimate the error rate.

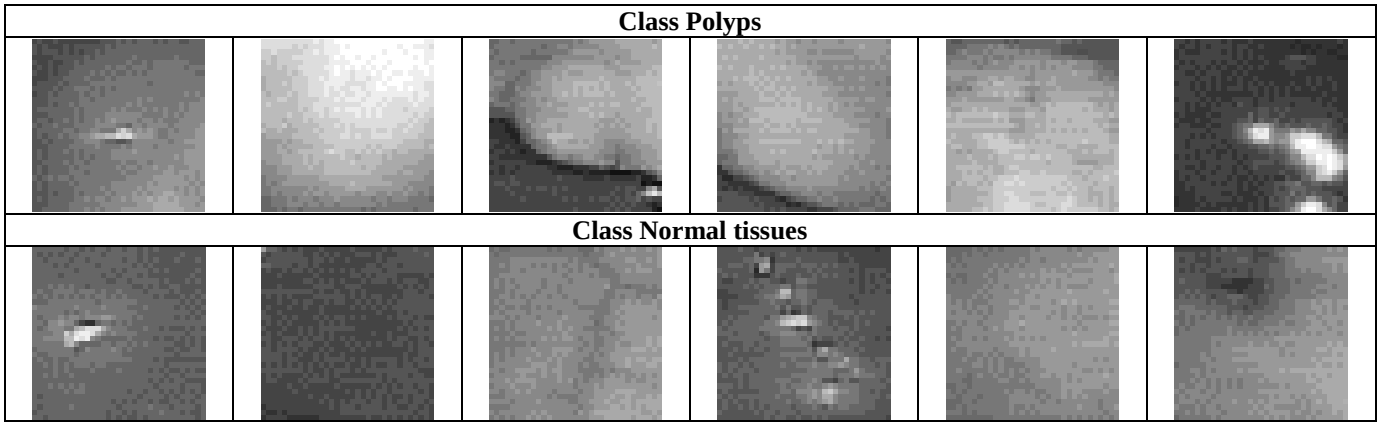

FIG. 1. SOME EXEMPLARY IMAGES

The tool ***Decision Master***® (Perner 1994-2017) is a data mining tool based on decision tree induction. It contains binary and n-ary decision tree induction methods such as standard algorithm like C4.5, ID3 and n-ary decision-tree induction methods developed by the Institute of Computer Vision and applied Computer Sciences ***IBaI*** (Perner & Trautzsch 1998).It is a commercial tool now and sold worldwide. It allows comparing the learnt models based on standard algorithms and special developed algorithms. N-ary trees get usually more compact than binary trees. The explanation capability of the trees is then better than for binary trees. However, it cannot be said from scratch which decision tree induction method is best based on the error rate for the desired data set. Therefore, it should be easy for the user to check out several decisions tree induction methods. The tool ***Decision Master***® allows that in an excellent manner. It has functions for dealing with missing values, erroneous values, and outliers such as the box-plot method and others. N-fold cross-validation or test-and-train can evaluate the learnt model. Several error rates are calculated such as the overall error rate, the class-specific error rate, and the classification quality (Perner et. al 2001). The tool has a nice user interface so that a non-computer expert can easy handle it. The other option is to integrate the software as OEM component in larger systems for example in E-commerce suites for on-line user profiling or for learning other information from the trace of the online-user (Perner & Fiss 2002).

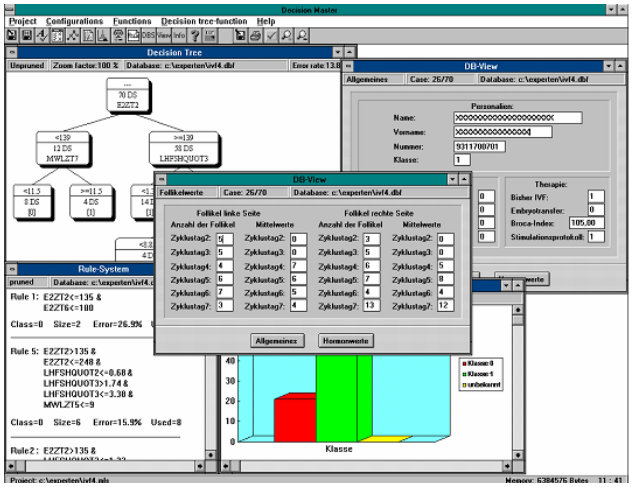

FIG. 2 SCREENSHOT OF THE DATA-MINING TOOL DECISION MASTER®

VI. 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.

Figure 3 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 4 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

(Perner et. al 2002).

In the first test (test_1), 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 (test_2), 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 COO-1 has an error rate of 3.33% and COO-2 has an error rate of 10% (see Figure 6).

S	**Polyp**	**Polyp**	**Polyp**	**Normal tissue**	**Normal tissue**	**Normal tissue**
Original image						
1						
2						
3						
4						
5						
6						

FIG. 3. THE IMAGES f(x,y,t) WITH S=6

S	Polyp1	Polyp6	Polyp20	Normal tissue	Normal tissue	Normal tissue
Original						
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						

FIG. 4. THE IMAGES f(x,y,t) WITH S=12

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 COO-1 has an error rate of 13.37% and COO-2 has an error rate of 18.89% (see Figure 7).

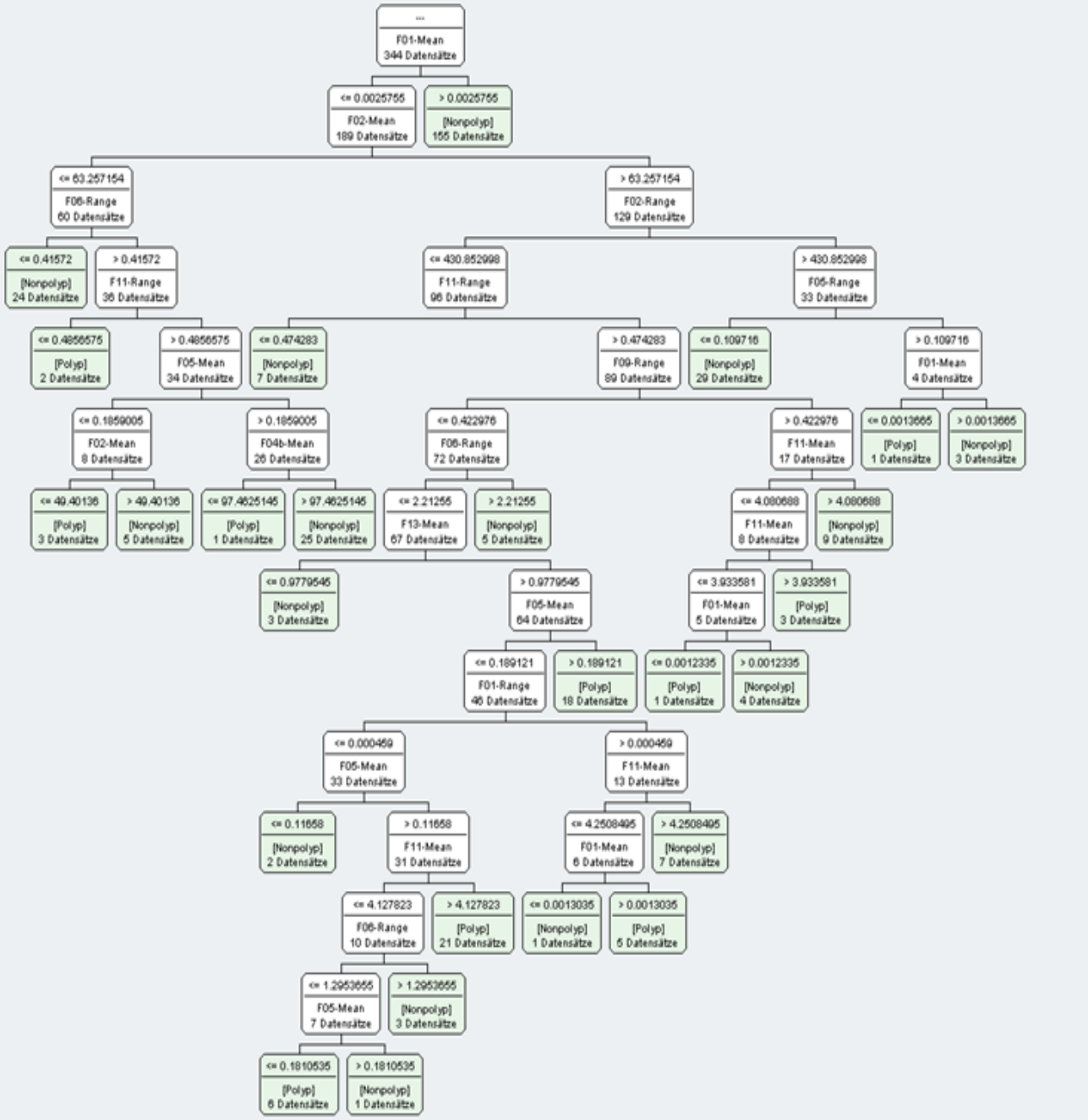

FIG. 5. DECISION TREE FOR COO FEATURE DESCRIPTOR

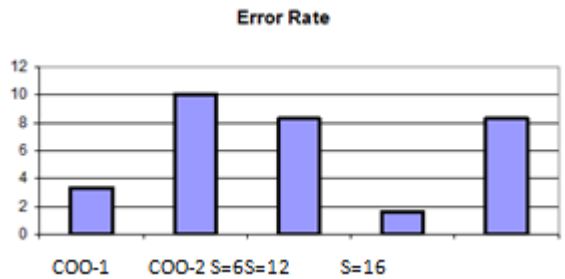

FIG. 6. ERROR RATE (IN PERCENT) FOR TEST 1

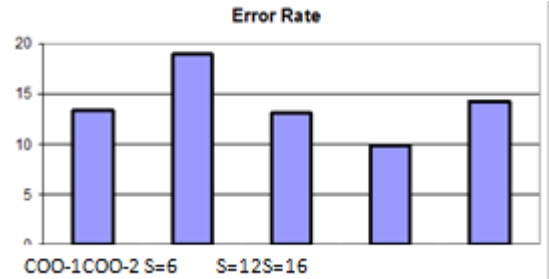

FIG. 7. ERROR RATE (IN PERCENT) FOR TEST 2

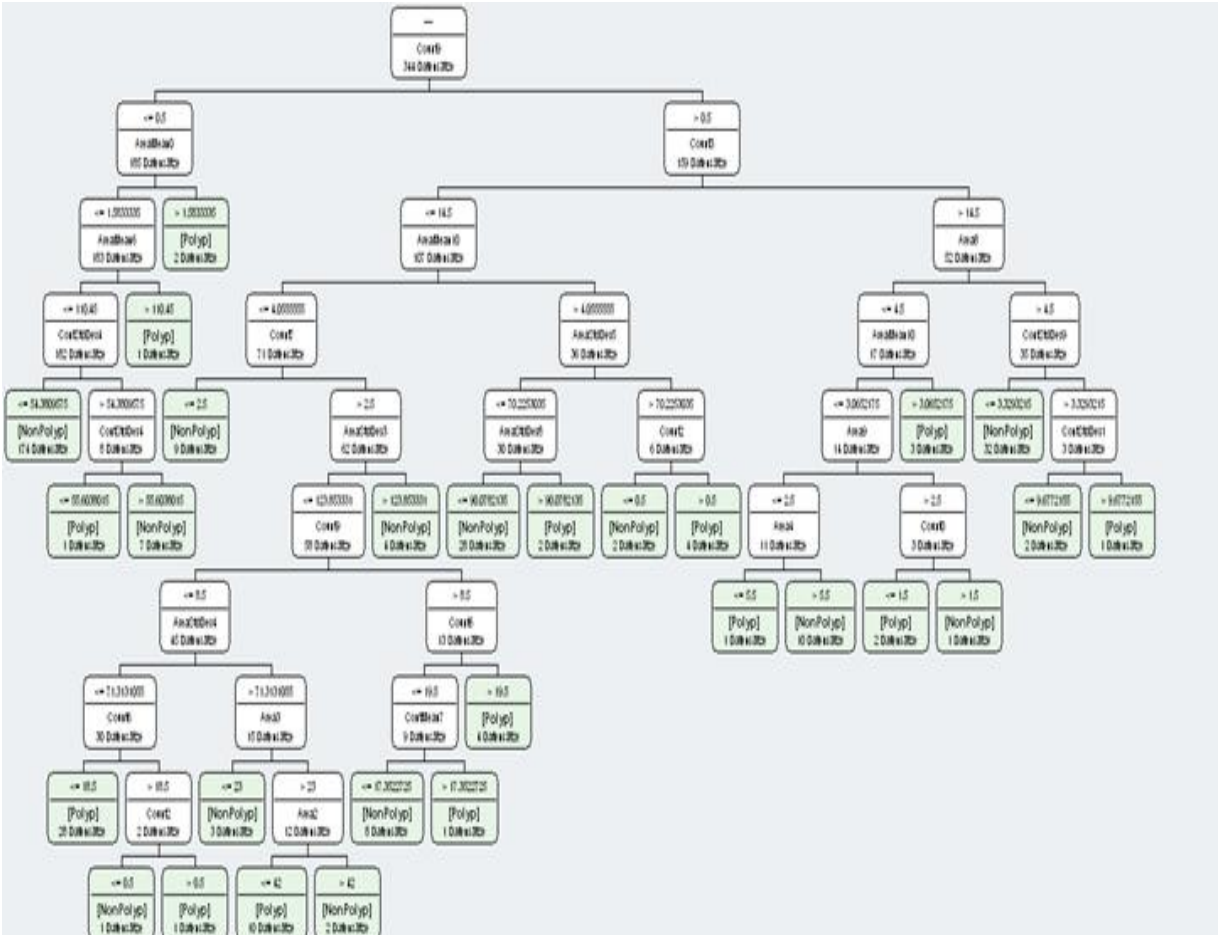

FIG. 8. DECISION TREE FOR TEXTURE FEATURES BASED ON RANDOM SETS

The resulting decision trees are shown in Figure 5 for COO feature descriptor and in Figure 8 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 COO texture descriptor and 22 features from 84 features for the texture descriptor based on random sets (see Table 2). The tree expands more in depth for the COO feature descriptor than for the texture descriptor based on random sets. The runtime of the program for the calculation of COO texture descriptor is 7-times longer than for the texture descriptor based on random sets (see Table 3).The runtime of the program for the calculation of COO-2 texture descriptor is not as long but the error rate is much higher than that for COO-1.

TABLE 3

RUN TIME

Runtime	COO-1	COO-2	Texture Descriptor based on Random Sets
	91.03s	83.22s	13.75s

VII. DISCUSSION

We have studied the behavior of the well-known co-occurrence texture descriptor to our novel random set texture descriptor. Both are statistical-based texture descriptors.

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

A further advantage of the texture descriptor based on random sets over COO 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 (see Table 2).

The choice of the number of slices *S* emerges to S=12 in all the applications we have done until now. 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-dimensionality). Besides that, our observations showed that the objects in the slice images converges to single points in case of S=16. If this happens then there is no information in the shape or contour anymore.

The Random-Set texture descriptor has a nice feature besides the co-occurrence texture descriptor. Semantic labels depending on the application can describe the texture and therefore the meaning of the texture is understandable by a human. In case of Figure 4 we can say for poly_1-image, it has objects in the higher slices. In case of poly_20-image, it is a homogenous texture since objects are distributed over all slices, and in case of normal tissue, the objects are in the middle of the slices. The semantic label helps the human to understand the texture and to talk about the texture in a common way. The medical texture objects such as for the polyp images and for cell images (Perner et. al 2002) are often not large objects. That limits the statistics we can use. We stayed on the first-order statistics. Higher-order statistics make no sense for small objects since the number of objects gets low and no sufficient static can be calculated.

The run-time of the random-set texture descriptor is seven times lower than as for the co-occurrence texture descriptor. This is a big advantage of the random-set texture descriptor over the co-occurrence texture descriptor. It helps to speed up the calculation of the image processing methods. The random-set texture descriptor can be given out a standard module for texture calculation. The input to the module is only the object points and the outputs of the module are the calculated features for the slices.

The decision tree induction method performs nicely on texture classification. The decision tree induction method is also a feature selector. Therefore, the method can as a learning method for the classification model as well as a feature selector. The number of features selected for COO 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 richer description of texture. Features from almost all slices are included in the decision.

The data set contains 283 normal tissue images and 61 polyp images. This is a two-class problem. The unequal number of the data in the two classes makes our problem to an unbalanced data set problem. The polyps in the images were identified and selected by a "well-trained" medical expert. The 283 normal images consist of dark regions and reflection. For the full unequally distributed data set, we achieved an error rate of 9.88% based on cross-validation. We achieved an error rate of 1.67% by cross-validation when we created a data set with equally distributed data in each class. To sample the data into two equally distributed data sets is our strategy to deal with the unbalanced data set problem for decision tree induction.

VIII. CONCLUSION

Many texture descriptors are known from the literature (Rao 1990). The most used texture descriptor is the texture descriptor based on the co-occurrence matrix. We proposed a texture descriptor based on Random sets (Perner et. al 2002a). In this paper, we compared both texture descriptors based on a medical-image data set for colon examination. The image should be classified into normal tissue images and into polyp images. We choose a medical application since the appearance of many medical objects can often be nicely described by texture. We learnt a classifier model based on decision tree induction. Then we compared the classification results for both texture descriptors.

We have found that the texture descriptor based on Random sets outperforms the co-occurrence texture descriptor based on the error rate, tree properties and the runtime. Co-occurrence 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 the co-occurrence texture descriptor has limited description power since the error rate is much higher than that for the texture descriptor based on random sets. One reason might be that the medical objects are not so large and the higher-order statistics fail due to the limited number of pixels. The run-time of the Random-set texture descriptor is seven times lower than as for the co-occurrence texture descriptor. This is a big advantage of the Random-set texture descriptor over the co-occurrence texture descriptor since the large computation time of image analysis algorithm is still a problem. The Random-set texture descriptor can form a software module that can be used for different applications and different sizes of the objects.

In addition, the texture descriptor based on Random sets has semantic meanings. An expert can understand the properties of the texture when looking at the slices produced during the calculation of the texture features. Therefore, the different appearances in the slices can be labeled by semantic terms that would give us explanation capability of the different textures.

The unbalanced data set problem as it often appears for medical data sets is handled in our study by sampling two equally distributed data sets together for the two-class problem. If we use this data set we can achieve a higher accuracy for the classification for both texture descriptors but still the Random-set texture descriptor outperforms the co-occurrence matrix-texture descriptor.

ACKNOWLEDGMENT

This work has been sponsored under the grant title “Study of the Cognitive Aspects of Human Vision” CogVision under the grant number IS 2012-4.

REFERENCES

[1] Rao A. R (1990).: A Taxonomy for Texture Description and Identification, Springer Verlag, Berlin.

[2] Haralick, R.H., Shanmugam, K., Dingstein, I. (1973), Textural Features for Image Classification. IEEE Transactions on Systems, Man and Cybernetics 3(6), 610-621.Perner P., Perner H., Müller B. (2002a), Mining Knowledge for Hep-2 Cell Image Classification, Journal Artificial Intelligence in Medicine (26), 161-173.

[3] M. H. Bharati, J. J. Liu, and J. F. MacGregor (June 2004), “Image texture analysis: methods and comparisons,” Chemometrics and Intelligent Laboratory Systems, vol. 72, pp. 57-71.

[4] G. Castellano, L. Bonilha, L. M. Li, and F. Cendes (Dec. 2004), “Texture analysis of medical images,” Clinical Radiology, vol. 59, no. 2, pp. 1061-1069.

[5] J. G. Zhang and T. N. Tan (Mar. 2002), “Brief review of invariant texture analysis methods,” Pattern Recognition, vol. 35, pp. 735-747.

[6] L. M. Kaplan (Nov. 1999), “Extended fractal analysis for texture classification and segmentation,” IEEE Trans. Image Processing, vol. 8, no. 11, pp. 1572-1585.

[7] T. M. Nguyen and Q. M. J. Wu (Feb. 2012), “Gaussian-mixture-model-based spatial neighborhood relationships for pixel labeling problem,” IEEE Trans. Systems, Man, and Cybernetics Part B-Cybernetics, vol. 42, no. 1, pp. 193-202.

[8] J.-L. Chen and A. Kundu (1993), “Automatic unsupervised texture segmentation using hidden Markov model,” in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Minneapolis, Minnesota, Apr. 27-30, pp. 21-24.

[9] J. Li, A. Najmi, and R. M. Gray (Feb. 2000), “Image classification by a two-dimensional hidden Markov model,” IEEE Trans. Signal Processing, vol. 48, no. 2, pp. 517-533.

[10] H. Dreyer and W. Sauer, Prozeßanalyse – ElementarestochastischeMethoden, VEB VerlagTechnik Berlin, 1982

[11] S. Krishnamachari and R. Chellappa (Feb. 1997), “Multiresolution Gauss-Markov random field models for texture segmentation,” IEEE Trans. Image Processing, vol. 6, no. 2, pp. 251-267.

[12] R. Haralick (May 1979), Statistical and Structural Approaches to Texture, IEPEREO, C EETDHINEG S OF VOL. 67, NO. 5, p. 786-804

[13] Luc van Gool, P. Deweale, A Oosterlink (1985), Survey – Texture Analysis Anno 1983, Computer Vision, Graphics, and Image Processing, 29, p. 336-357.

[14] Din-Chang Tseng, Member, IACSIT and Ruei-Lung Chen (June 2015), Multiscale Texture Segmentation Using Contextual Hidden Markov Tree Models, International Journal of Machine Learning and Computing, Vol. 5, No. 3, p. 198- 205.

[15] Wesley NunesGon_calves, Bruno Brandoli Machado, and Odemir Martinez Bruno (Febr. 2014), Texture descriptor combining fractal dimension and artificial crawlers, Physica A: Statistical Mechanics and its Applications, Volume 395, Pages 358–370

[16] R. Mukundan (2014), Local Tchebycheff Moments for Texture Analysis, In: G.A. Papakostas, Moments and Moment Invariants - Theory and Applications, GCSR Vol. 1, Science Gate Publishing, p. 127- 142.

[17] YuhuiQuan , Yong Xu, , Yuping Sun (April 2014), A distinct and compact texture descriptor, Image and Vision Computing, Volume 32, Issue 4, Pages 250–259

[18] J.-F. Aujol, G. Gilboa, T. Chan, St. Osher (April 2006), Structure-Texture Image Decomposition - Modeling, Algorithms, and Parameter Selection, International Journal of Computer Vision, Volume 67, Issue 1, pp 111-136

[19] I. Champion, Chr. Germain, J. P. Da Costa, A. Alborini, and P. Dubois-Fernandez (January 2014), Retrieval of Forest Stand Age From SAR Image, Texture for Varying Distance and Orientation Values of the Gray Level Co-Occurrence Matrix, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, VOL. 11, NO. 1, p. 5-9.

[20] Ch. Dharmagunawardhana, S. Mahmoodia, M. Bennettb, M. Niranjana (Nov. 2014), Gaussian Markov random field based improved texture descriptor for image segmentation, Image and Vision Computing, Volume 32, Issue 11, Pages 884–895.

[21] H. Madzin, R. Zainuddin, and Nur-Sabirin Mohamed (Nov. 2014), Analysis of Visual Features in Local Descriptor for Multi-Modality Medical Image, The International Arab Journal of Information Technology, Vol. 11, No. 5, p. 468- 475.

[22] N.Palanivel, P.Keerthika, K.Yazhini, P.Thamizhini (April 2015), Texture Analysis using Markov process with Bayesian Approach, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 4, p. 101-104.

[23] J. Massich, F. Meriaudeau, M. Sent´ıs, S. Ganau, E. P´erez, D. Puig, R. Mart´i, A. Oliver, and J. Mart´I (2014), SIFT Texture Description for Understanding Breast Ultrasound Images In: H. Fujita, T. Hara, and C. Muramatsu (Eds.): IWDM 2014, LNCS 8539, pp. 681–688, 2014, Springer International Publishing Switzerland.

[24] Tiecheng Song, Hongliang Li, Senior Member, IEEE, FanmanMeng, Qingbo Wu, Bing Luo, Bing Zeng (2014), Noise-Robust Texture Description Using Local Contrast Patterns via Global Measures, Signal Processing Letters, IE Volume:21 Issue:1, p. 93 - 96

[25] B. Zhang, Tuan D. Pham (2010), Multiple Features Based Two-stage Hybrid Classifier Ensembles for Subcellular Phenotype Images Classification, International Journal of Biometrics and Bioinformatics, (IJBB), Volume (4): Issue (5) 176-193.

[26] Tuan D. Pham, (January 10, 2014) Automated identification of mitochondrial regions in complex intracellular space by texture analysis, Proc. SPIE 9069, Fifth International Conference on Graphic and Image Processing (ICGIP 2013), 90690G; doi:10.1117/12.2050102

[27] J. V. Marcos,, R. Nava, G. Cristobal, R. Redondo, B. Escalante-Ramrez, G. Bueno, O. Deniz, A. Gonzalez-Porto, C. Pardo, F. Chung, T. Rodriguez (2015), Automated pollen identification using microscopic imaging and texture analysis, Micron, Volume 68, January 2015, Pages 36–46

[28] J. Olveres, R. Nava, E. Moya-Albor, B. Escalante-Ramirez, J. Brieva, G. Cristobal and E. Vallejo (2015), Texture descriptor approaches to level set segmentation in medical images, Optics, Photonics, and Digital Technologies for Multimedia Applications III, edited by Peter Schelkens, TouradjEbrahimi, Gabriel Cristóbal, FrédéricTruchetet, PasiSaarikko, Proc. of SPIE Vol. 9138, p. 1-12

[29] L. Cai, X. Wang, Y. Wang, Y. Guo, J. Yu and Y. Wang (2015), Robust phase-based texture descriptor for classification of breast ultrasound images, BioMedical Engineering OnLine (2015) 14:26, p. 1-21

[30] Da-Chuan Cheng, Wen-Chien Ting, Yung-Fu Chen, Qin Pu, Xiaoyi Jiang (2008), Colorectal Polyps Detection Using Texture Features and Support Vector Machine, In: P. Perner, Advances in Mass Data Analysis of Images and Signals in Medicine, Biotechnology, Chemistry and Food Industry Lecture Notes in Computer Science Volume 5108, 2008, pp 62-72

[31] P. Perner (1998), Classification of HEp-2 Cells Using Fluorescent Image Analysis and Data Mining,14th Intern. Conference on Pattern Recognition, Brisbane Austr., IEEE Computer Society Press 1998, Vol. II, pp. 1677-1679

[32] Matheron, G. (1975), Random Sets and Integral Geometry. J. Wiley&Sons, New York, London (1975)

[33] Stoyan, D., Kendall, W.S., Mecke, J. (1987), Stochastic Geometry and Its Applications. AkademieVerlag (1987)

[34] Perner P. (1994-2017), Data Mining Tool *Decision Master*, ibai-solutions www.ibai-solutions.de

[35] Klette, R., Zamperoni, P. (1996), Handbook of image processing operators, Chichester; New York: Wiley, 1996.

[36] Perner P. (2002b), Data Mining on Multimedia Data, lncs 2558, Springer Verlag, 2002.

[37] Perner P. and Trautzsch S. (1998), Multinterval Discretization for Decision Tree Learning, In: Advances in Pattern Recognition, A. Amin, D. Dori, P. Pudil, and H. Freeman (Eds.), Springer Verlag 1998, LNCS 1451, pp. 475-482.

[38] P. Perner, U. Zscherpel, C. Jacobsen (2001), A Comparision between Neural Networks and Decision Trees based on Data from Industrial Radiographic Testing, Pattern Recognition Letters 22 (2001), pp. 47-54.