Image analysis based grading of bladder carcinoma. Comparison of object, texture and graph based methods and their reproducibility

Heung-Kook Choi[a,e], Torsten Jarkrans[a], Ewert Bengtsson[a], JanosVasko[b], Kenneth Wester[c], Per-Uno Malmström[d] and Christer Busch[c]
[a]*Centre for Image Analysis, Uppsala University, Uppsala, Sweden*
[b]*Department of Pathology, Umeå University, Umeå, Sweden*
[c]*Department of Pathology,* [d]*Department of Urology, University Hospital, Uppsala, Sweden*
[e]*School of Information and Computer Science, Inje University, Kim-Hae, Korea*

Received 17 September 1996
Revised 21 May 1997

Abstract. The possibility that computerized image analysis could increase the reproducibility of grading of bladder carcinoma as compared to conventional subjective grading made by pathologists was investigated. Object, texture and graph based analysis were carried out from Feulgen stained histological tissue sections. The object based features were extracted from gray scale images, binary images obtained by thresholding the nuclei and several other images derived through image processing operations. The textural features were based on the spatial gray-tone co-occurrence probability matrices and the graph based features were extracted from the minimum spanning trees connecting all nuclei. The large numbers of extracted features were evaluated in relation to subjective grading and to factors related to prognosis using multivariate statistical methods and multilayer backpropagation neural networks. All the methods were originally developed and tested on material from one patient and then tested for reproducibility on entirely different patient material. The results indicate reasonably good reproducibility for the best sets of features. In addition, image analysis based grading showed almost identical correlation to mitotic density and expression of p53 protein as subjective grading. It should thus be possible to use this kind of image analysis as a prognostic tool for bladder carcinoma.

Keywords: Image analysis, feature measurements, grading of bladder carcinoma, reproducibility

1. Introduction

A few years ago we started a project with the goal of finding a straightforward, reproducible and objective computerized grading system for bladder tumors. Some preliminary results have been reported from this work [8,9,19]. The very first system for such grading was introduced by Broders [4] in 1922. In 1966 Bergkvist et al. [2] proposed that four grades be used and our material is graded according to a modification of this system published by Malmström et al. [30,36]. The grading is based on subjective judgement of a range of structural changes, mainly changes of harmony in the pattern found by the cellular elements in the stained histological section [5]. In 1973 the World Health

0921-8912/97/$8.00

Organization (WHO) introduced a standard according to which the urothelial tumors are classified into three grades. Since the grading is clinically important but the reproducibility of the visual grading may be questioned [33] it is worthwhile looking for alternative, more objective ways of grading. Computerized image analysis is an obvious possibility.

Quantitative, computer assisted, analysis of tumor tissues started in the late sixties [39]. Features that are not readily apparent to the human eye such as high order texture descriptors can also be significant. Since much of the information used to grade the tissue comes from the relations between the cells, i.e., an assessment of the degree of order of the disorder of the tissue organization, it is interesting to try to describe these relations. Prewitt proposed that graphical analysis be based as a mathematical tool for this [40]. Since then several other authors have followed up the idea and developed graph analysis algorithms for tissue characterization [13,22–24,31]. In particular the Minimum Spanning Tree (MST) has proven useful to describe the tissue architecture.

The aim of this study was (a) to utilize retrospective patient material to select objects, texture and graph based features, single and in combination, which give optimal correlation to subjective grading, (b) to use these features on a second prospective patient material in order to test the correlation of these, single and in combination, to subjective grading, to mitotic density and to p53 protein expression, and (c) to test the reproducibility of the image analysis based grading when implemented on different scanner and computer systems and used on new patient material.

2. Materials

The study was based on two entirely different patient materials. These were collected at different time periods, at different hospitals and analyzed with different computer systems. The first material was used for developing and initially testing the methods. The second material was only used for studying how reproducible the methods were when applied under new conditions.

Material I: All patients with newly detected transitional cell carcinoma of the urinary bladder who were seen at Uppsala University Hospital during the period 1975–1978, were studied retrospectively. The total series comprised 230 patients with an observation time ranging from 5 to 9 years (mean 6.5 years) [30]. In 197/230 cases there was sufficient histological material left from the first biopsy to allow us to conduct computerized image analysis.

Material II: Bladder cancer biopsy samples were collected from 12 hospitals in Norway and Sweden participating in a clinical multicenter study [29]. The total series was composed of 261 patients who were included from 1987 to 1993. From these 100 cases remained those with papillary tumor and sufficient histological material left from the primary biopsy to allow us to conduct computerized image analysis. From these cases pathologists chose two regions of interest in each specimen so that we obtained 200 gray-scale images.

3. Methods

3.1. The subjective grading

Subjective grading is used to divide the material into four groups: Grade 1, 2A, 2B and 3. Grade 1 (Fig. 1a) is characterized by slight nuclear atypia and enlargement, it usually includes nuclei with a slight tendency towards variation in the orientation of the major axis. The overall impression is that

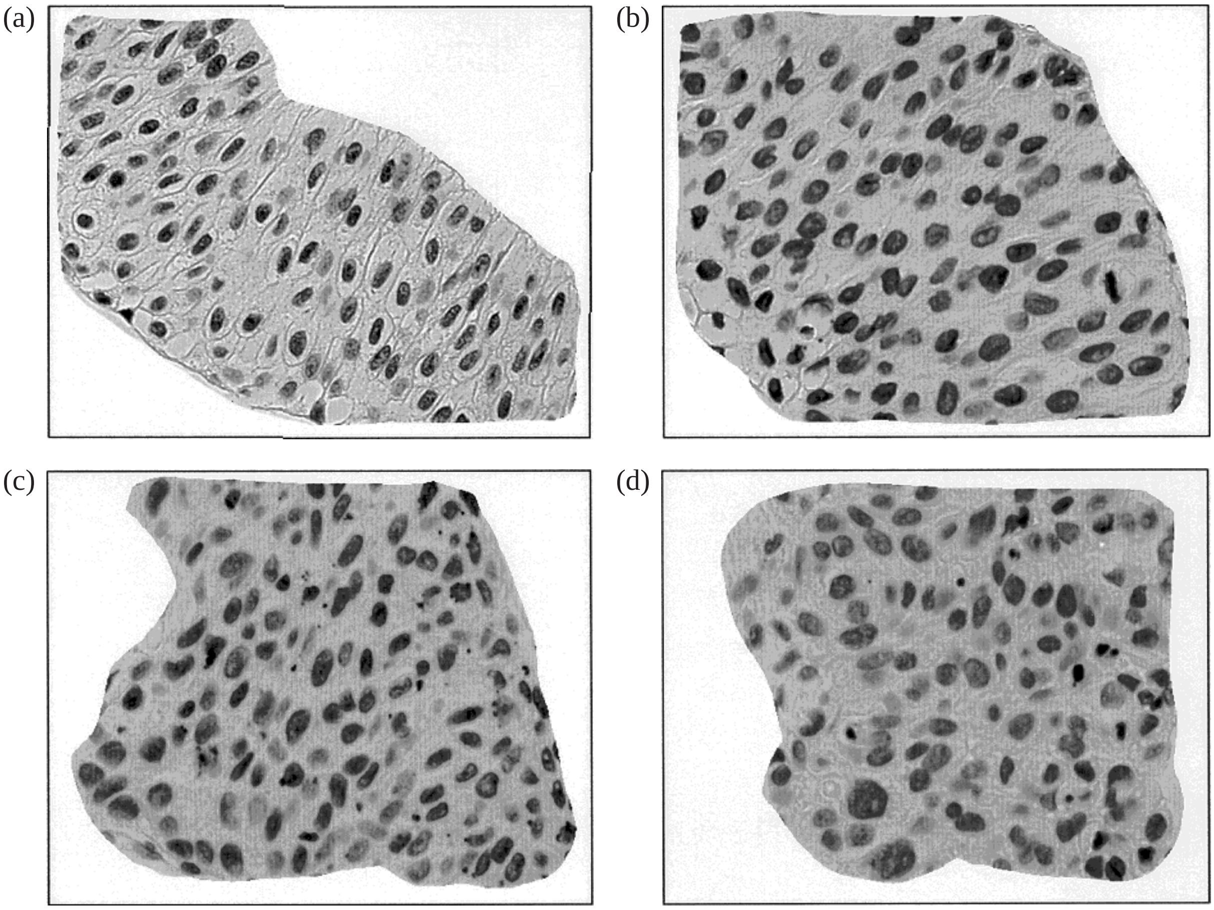


Fig. 1. Digital microscopic images (40×) of the bladder carcinoma cell stained by Feulgen. In each grade the region of interest was chosen by pathologists: (a) Grade 1, (b) Grade 2A, (c) Grade 2B and (d) Grade 3.

of an ordered pattern with very little variation. In Grade 2A (Fig. 1b) the enlargement of the nuclei is more easily recognized and the variation in nuclear staining density more pronounced. The cellular pattern still gives a predominantly ordered impression but with more obvious disorder because of the variation in the orientation of the major nuclear axis, as well as variation in the nuclear size and shape, internuclear distances and chromatine features. In Grade 2B (Fig. 1c) the general impression is one of disorder, but with focal areas of order. The disorder is due to more pronounced deviations in each of the individual features listed above. Grade 3 (Fig. 1d) is the case with an even more pronounced impression of disorder, often based on extreme variations in each of the features described.

3.2. Image acquisition

Digitized images were obtained from 5 μm thick paraffin sections that were Feulgen stained (acid hydrolysis in 5 M HCL at 22°C for 60 min, stained with Schiff's regent at 22°C for 120 min) [43]. A microscope with an interference filter (IF) 550 nm and an objective magnification of 40× made the size of the pixels about 0.5 μm [46]. The image was corrected for shading and quantisized into

8 bits of gray scale. The regions of interest that were analyzed were chosen by pathologists to be those most representative for the subjective grading.

All image analysis was done on software systems developed at the authors laboratories over the last 20 years.

For materials I all image processing was carried out on an EPSILON workstation from IMTEC using the EGO software system [1,32]. A standard monochrome video camera and a frame-grabber giving a data matrix of 512 × 410 pixels was used.

For materials II: A software package, Signifier, from Diascan AB in Uppsala was used on a Silicon Graphics Indy workstation for the image handling. The gray scale image in this case was obtained through the frame-grabber built into the workstation connected to a standard monochrome video-camera and had a matrix size of 756 × 572 pixels.

3.3. Image analysis strategies

Three different, rather independent, strategies have been applied when trying to express the characteristics of the tissue which is the basis for the visual grading. These strategies have previously been described [8,9,19]. In order to make this paper more readable the strategies are briefly outlined here; we refer to the original papers for details. In some cases they have been extended and modified as described below.

3.4. Object based image analysis (*Object*)

Since the subjective visual image analysis of the histological sections can be expressed in terms of order/disorder as far as the distance between the objects, the variation in the gray scale between objects and the relative size and orientation of objects in the image are concerned, it seems reasonable to assume that a computerized image analysis procedure might express the relevant image properties. For this purpose we have developed a fully automatic image analysis procedure operating on the gray scale image defined by the region of interest. The gray scale image is first segmented into a binary image with objects (nuclei) and background. The binary image is then filtered to further improve the result from the thresholding and a set of derived images is created (*skeleton, distance, zoi, gradient magnitude and gradient direction image*).

3.4.1. Feature extraction

Features were extracted from the gray scale image, the binary image and the derived images. Four sets of features were calculated: *simple object features, zoi object features, neighborhood object features, and non-object based features* [19].

Rather than using the directly derived features as such we used the frequency histogram of each of the primary features from all four sets (simple, zoi, neighborhood, non-object based) obtained from all objects, e.g., the distribution of the size of the objects. From such a distribution we pick the *median* (p50), 15% *percentile* (p15), 85% *percentile* (p85), as shown in Fig. 2. The purpose of this procedure is to define features on the image (and tissue) level rather than the individual object level and to overcome the problem of extreme values in the distributions. We then defined the coefficients of percentile variation, CPV, as Eq. (1):

$$\mathrm{CPV}_{8515} = \frac{(\mathrm{p}85 - \mathrm{p}15)}{(\mathrm{p}85 + \mathrm{p}15)}. \tag{1}$$

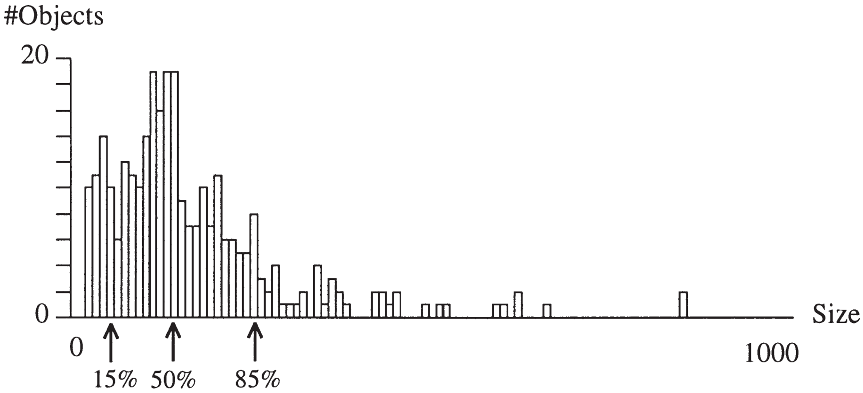


Fig. 2. The frequency histogram of object size for a Grade 1 image. The three percentiles that are used as image features are marked.

The purpose of the CPVs is to create image features that are highly sensitive to variations and much less sensitive to absolute measurements from the image, of, for example, the size of the objects.

Numerous features were originally calculated and subjected to analysis by multivariate methods. From this analysis a small set of features proved to be the best in expressing the grade of the training sets. The five features currently used are variations in all edge directions in the image, the standard deviation of the gray scale of objects (CPV_{5015}), the size of the objects, the standard deviation of distances inside objects (CPV_{8515}) and the standard deviation of the major axis for neighboring objects (CPV_{8515}).

3.5. *Textural image analysis* (*Texture*)

Since an important basis for the visual grading of the tissue sections is the general appearance of the section we felt that a texture analysis of the section would be useful. Thus we do not analyze the texture of the individual nuclei as is done in many other previously published studies. We perform a texture analysis on the entire selected region of interest of the tissue section without previous segmentation. For the region of interest, a gray scale co-occurrence probability matrix is generated. This matrix contains an estimate of the joint probability density function for pixels separated by a particular number of rows, diagonals or columns. If the co-occurrence matrix was calculated directly from the original image it would become very large and highly dependent on both the illumination level and contrast setting. This problem was reduced through a histogram equalization operation. In order to be able to split pixels that came within one bin in the original histogram into two bins in the new histogram, a randomized selection process was necessary. This was done by adding uniformly distributed noise to the image. The original 8 bits gray scale was extended to 12 bits through the application of the pseudo-random noise function. This technique normalizes all images to the same uniform gray level frequency distribution. The co-occurrence matrix $P(i,j|d,\theta)$ contains the probability that the gray level i occurs at a distance d in direction θ from a pixel with value j in the region of interest in the image. Whereas the minor axes of bladder carcinoma nuclei are about 8 μm, the major axes are about 13 μm, and the internuclear distances are typically a few nuclear diameters so we should study step-lengths in this range. We created 12 matrices: the first four matrices were for the angles, the second four matrices were from the maximum, minimum, mean and the difference between maximum and minimum for measurements of the four angles, and the third four matrices were for extracting the maximum gray value, minimum gray value, mean gray value and the difference value between maximum and minimum among the four different angles. Then each matrix was normalized. These resulting measures were made invariant to any image rotations.

3.5.1. Calculation of features

The co-occurrence matrices as such contain too many values to be used directly. Several secondary features, mainly weighted moments that condense the information in the matrices, were calculated. Haralick et al. [18] proposed a set of moments and Pressman [38] suggested additional ones. Using these 28 moments for the maximum, minimum, mean and difference matrices, we got a total of 224 features for each step-length. The most significant features are angular second moment, inverse difference moment, diagonal moment, mean variation, correlation, difference average, sum average, sum squares of variance of horizontal direction, diagonal variance, sum variance, difference variance and triangular symmetry. The inverse difference moment measures regularity. The correlation provides a correlation coefficient of the rows, diagonals and columns of the normalized co-occurrence matrix. When the co-occurrence matrix contains many values near the diagonal the contrast is small, but the inverse difference moment is large [10].

3.6. Graph based analysis (*Graph*)

For this approach we needed to segment the tissue sections to find the nuclei. The same histogram based global thresholding used for the object based image analysis was used for this purpose. It is based on the fact that cell nuclei appear as dark objects on a lighter background. We found the threshold by searching for minima in the first derivative of the histogram. The derivative transforms inflection points to minima. The resulting binary image (Fig. 3b) showed the nuclei but also many other spurious objects and holes in the nuclei. Another problem was caused by touching or overlapping cells. The object masks were first cleaned by filling the holes in the objects. Small objects, defined as having an area less than 70 pixels or a maximum width of less than 9 pixels, were deleted. The width was obtained by a chamfer 3–4 distance transformation [3]. The distance transform image (Fig. 3c) was also used as a starting point for separating touching objects. The watershed segmentation algorithm [47] can separate irregular "blobs" into more regular sections. If all watershed regions are considered to be separate objects, excessive splitting will occur. To avoid this we empirically determined that a "waistline" between blobs with a width of less than 90% of the diameter of the smallest blob side should result in a cut. For this cut the watershed line was replaced with the shortest straight line [35]. By applying the watershed segmentation recursively, multiple overlaps could be handled, at least for simple cases. The resulting binary image is shown in Fig. 3d.

3.6.1. Minimum spanning trees

A minimum spanning tree connecting the centroids of all nuclei in a tissue section can be used as a basis for extracting features describing the tissue architecture. Thus one first creates a linear graph connecting the centre points of the nearest objects. These points form the vertices of the graph. The edges are assigned the Euclidean distance between the corresponding vertices as weights. This graph is then reduced to a MST using Kruskal's algorithm [25]. In this algorithm, edges are first sorted in ascending order of weight and then placed in a list structure. This algorithm was implemented here using a heapsorted priority queue [7]. As the list is traversed, if an edge is found connecting a vertex that has not been visited before, the edge is included in the tree. When an edge connects two vertices belonging to different clusters, the two clusters are merged, with the new cluster being absorbed into

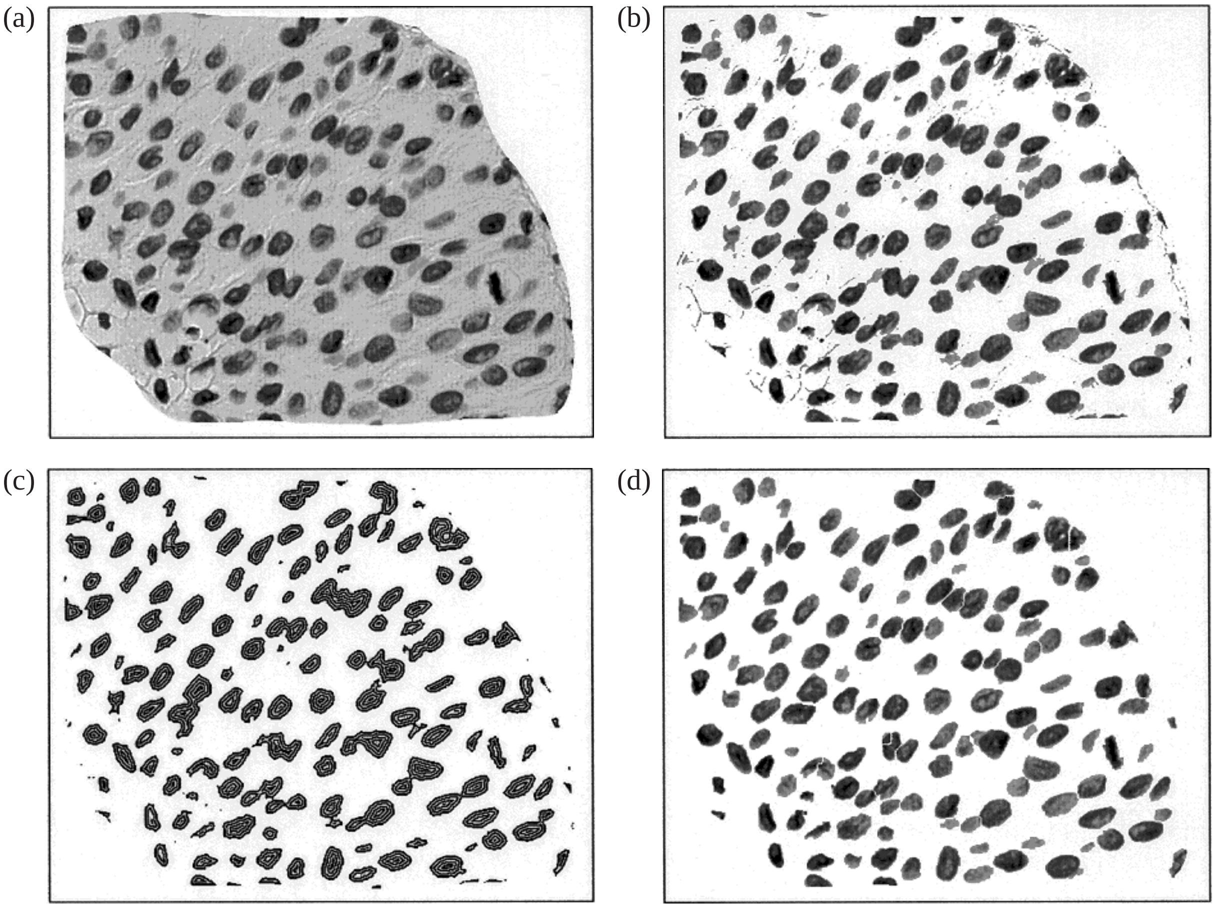
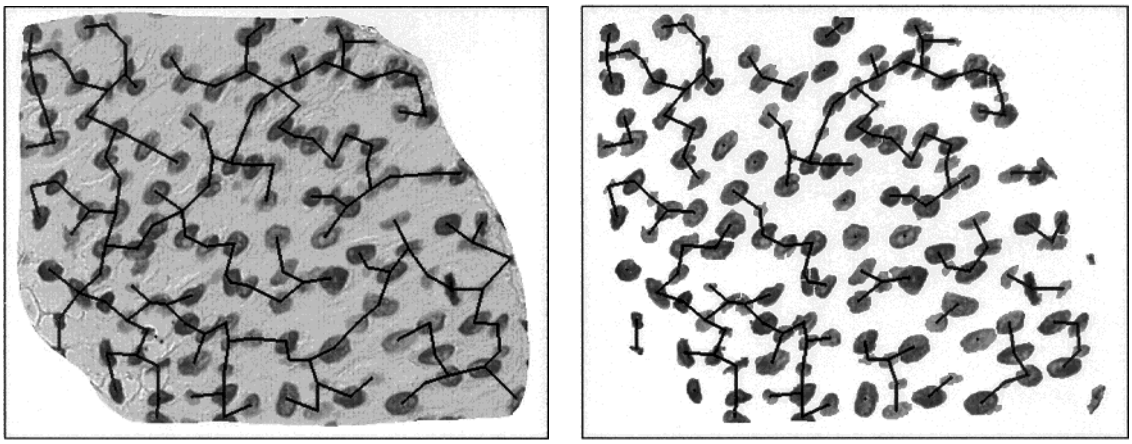


Fig. 3. (a) The Grade 2A image. (b) Cell nucleus segmentation of histogram-based global thresholding. (c) Using of a chamfer 3–4 distance transformation. (d) The overlapping nuclei separated by watershed algorithm.

the older ones. The tree will be completed when only one cluster remains, i.e., when all the vertices have been visited. For N vertices, the complete tree will have $(N - 1)$ edges. The tree obtained will be the minimum spanning tree (MST), i.e., the graph whose weight-sum is minimum among all spanning tree graphs. A practical example is shown in Fig. 4a. In the MST the vertices (nuclei) were clustered by cutting links longer than a certain threshold value. Reasonable values for this threshold were obtained by studying the image geometry: two slightly overlapping nuclei would have an edge distance of about 6 μm and a distance of more than 19 μm is rather unusual. Several distances in this range were used as possible thresholds for creating clusters. The threshold which yielded features with the best discriminating power was finally used. Figure 4b depicts the clusters of segmented nuclei.

3.6.2. Feature extraction

A number of geometric and densitometric features were measured for the clustered objects in the MST. We defined 24 different features. The best of these, as determined by their discriminatory power, were: the average polarity of the vertices, i.e., the average number of edges meeting at each vertex, the variation of local orientation of the graph, which was defined as the standard deviation of the directional indices measured in a 2×2 or 3×3 neighborhood, the standard deviation of the cluster



(a) (b)

Fig. 4. (a) The minimum spanning tree of a tissue section with 153 vertices superimposed on the gray scale image. (b) The Euclidean distance between vertices has been thresholded at 16 μm to create the 31 clusters.

area, the total number of clusters which contain only a single nucleus divided by the total number of nuclei, and finally, the variation in the average nuclear size, which was defined as the standard deviation of the average nuclear areas in each cluster, taken over all clusters.

3.7. *Mitotic density* (*MD*)

Sections stained with our modification of Gomori's silver methenamine method [6,15] were screened for areas with the highest mitotic activity. A rectangular part of the visual field, about 0.055 mm^2, was depicted on the computer screen (Fig. 5). A representative area of the tumour was encircled and the number of mitoses was counted visually (MA_f^n). This partial (A_p) and the number of nuclei in each A_f was counted manually (NA_p^n). The average of mitotic frequency (MF) of the five fields with the highest MF_f [46] was regarded as the MF of the tumour (MF_t), see Eq. (2):

$$MF_t = \frac{\sum\limits_{i=0}^{4} MA_f^n \times 1000}{\sum\limits_{i=0}^{4} (A_f/A_p) \times NA_p^n}. \tag{2}$$

3.8. *p53 protein expression*

Mutation of the p53 gene is one of the most commonly occurring genetic lesions associated with human cancer [16]. Wild type p53 protein acts as a tumour suppressor, either by arresting the cell in G1, allowing the cell additional time for DNA-repair, or by promoting apoptosis [28]. These reactions are triggered in response to genetic damage caused by irradiation or other carcinogenic agents [48]. Defective p53 protein is not functional. Mutations in the p53 gene alter the amino acid sequence in the DNA-binding domain of the protein and/or alter the three-dimensional configuration of the protein structure. The lack of functional p53 protein results in completion of the cell cycle, even in cells harbouring DNA-damages, and give rise to genetically defect daughter cells. The protein encoded

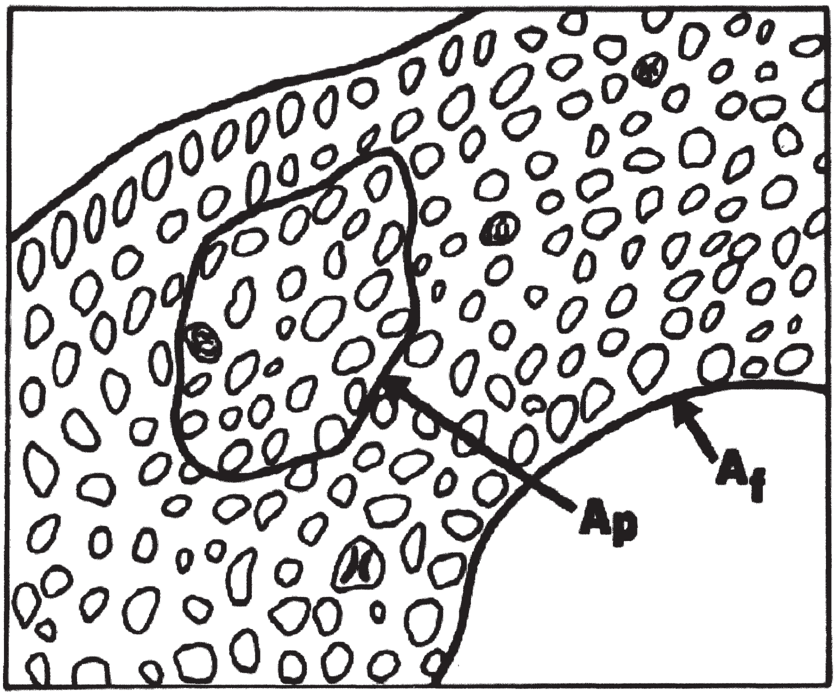


Fig. 5. The tumour epitherium (A_f) and a representative part of it (A_p) are marked. The number of mitoses in A_f is 4 ($MA_f^n = 4$).

by a mutated p53 gene exhibitis a dramatically increased half-life, compared to the wild-type protein [17], resulting in a stabilized protein localized in the nucleus. Furthermore, the possibility of detecting the stabilized p53 protein, using conventional immunohistochemistry (IHC) techniques, makes p53 a candidate prognostic marker in tumour pathology.

3.8.1. Evaluation of p53 protein expression

A scoring system based on both the extent and intensity was used in the evaluation of the p53 IHC. The extent was given a score from 1 to 3 ($1 < 25\%$, $2 = 25$–50% and $3 > 50\%$ of tumour cells stained). The intensity was judged as: 1 = negative, 2 = weak, 3 = moderate and 4 = strong staining. The product of the extent and intensity scores, resulting in a number between 1 and 12, were then divided into three classes as follows:

low or absent p53 expression: 1–2
moderate p53 expression: 3–4
high p53 expression: 5–12

The evaluation of the staining was performed in collaboration by two observers. Only nuclear staining was judged as positive.

3.9. Classifications

3.9.1. Statistical multivariate classification

To select the best set of features among all those calculated we applied a step-wise linear discriminant analysis using the BMDP [11] program package. In total, 26 features were selected for dividing the four groups into Grades 1, 2A, 2B and 3. Of these, five were object analysis features, fifteen were texture features and six were graph analysis features. We also tried to split the material into only two groups Grade 1 and 2A in the first group and Grade 2B and 3 in the second. For this 25 features were used. This program finds the optimal combination of features and the corresponding weights in the discriminant function [12,21,26]. The classification function is $g_i(w)$ for the object group i, see Eq. (3):

$$g_i(w) = \sum_j c_{ij} w_j + w_{i0}. \tag{3}$$

These functions are evaluated for each case which is assigned to the group for which the classification function has the largest value.

3.9.2. The classification of backpropagation neural networks

An input layer of 26 neurons represented the features which were normalized to have a value between 0 and 1. The connection weights w_{ij} and w_{jk} were both initialized to random values between 0.001 and 0.999. However, the neural network model of the backpropagation learning algorithm updates the values of the connection weights thereby reducing the error value. The relations between the input and the desired output are "learned" by the network through iterative adjustments of the connection weights such that the sum of the squared errors is minimized at the output layer for the training set [14]. The layer of 20 hidden neurons was included to represented possible interactions among the input variables. A sigmoid function was applied to each neuron in the hidden as well as in the output layers. The output layer of four neurons represented the class identity. The network learned to classify each training sample by adjusting the output neuron representing the correct class toward a value of 1, and the other towards 0. A gain term controls the adaptation rate and the convergence criterion during the network training was the achievement of an error level less than 10^{-6}. In order to learn successfully, we want to make the output of the net approach the desired output, that is, we want to continually reduce the value of this error function. The schematic architecture is shown in Fig. 6.

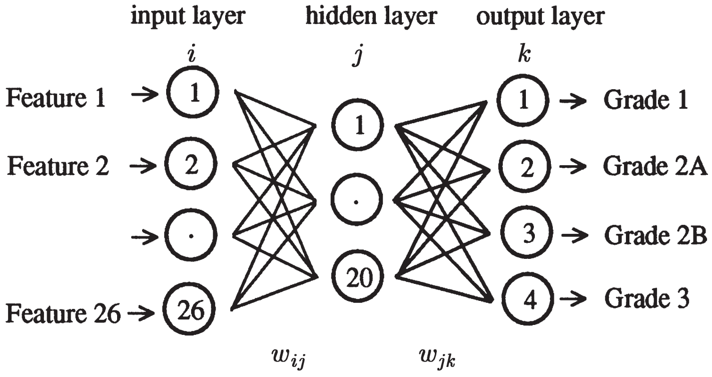


Fig. 6. Schematic diagram for a back propagation neural network. The w_{ij} was the weight from applied neuron i to j. The input layer consisted of 26 features. Each of these was connected to the hidden layer of 20 neurons, which was in turn connected to the output, and which was representative for the four-grading classification.

4. Results

In developing our computerized grading system we used visual grading as “the absolute truth”. Since visual grading is known to have low and variable reproducibility [34] it is important to verify the reproducibility and prognostic value of the resulting computer grading. The reproducibility was assessed in two ways, by subdivision of grades into two groups and four groups, respectively. In addition both subjective and computer grading were correlated to MD and expression of p53 protein, both known to be of prognostic significance.

4.1. Correlation of material I with patient survival

In a preliminary study early in this project we correlated the grouping obtained through conventional visual analysis and the grouping obtained through texture analysis to the survival time for material I as depicted in Fig. 7. The follow-up covered a period of 120 months. The texture classifier obtained 84.3% (kappa $=$ 0.787) agreement [8]. As can be seen from the graphs, the texture based grading seems to produce a more pronounced division regarding survival rate between groups 1 and 2A on the one hand and groups 2B and 3 on the other for the early survival times than does visual grading. These results need to be verified in the future when survival data becomes available for the independent, prospective material but they did encourage us to continue this work.

4.2. Correlation of the two different materials analyzed on different image scanner and computer systems

The two different materials were analyzed on two different image scanning and analysis systems. The training set from material I consisted of 197 images with a matrix size of 512 × 410 pixels and the test set from material II consisted of 200 images with the matrix size 756 × 572 pixels. We carried out a reproducibility test from the training set to the test set for both two groups and four groups. The results are presented in Table 1.

4.3. Correlation of two regions of interest in each specimen

Pathologists selected two regions of interest in each specimen from material II. The first group, consisting of 100 images (training set), was used for training, i.e., to create a linear discriminant

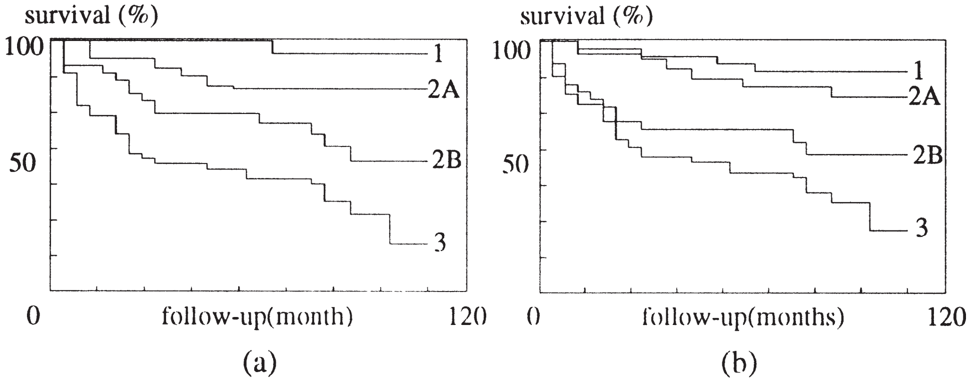


Fig. 7. The graphs were calculated according to the Kaplan–Meier method. (a) Survival related to subjective grade. (b) Survival related to texture analysis based grade.

Table 1
Reproducibility for the two independent materials and image analysis systems. Two different classification strategies and three feature extraction methods were analyzed. The percentages express the fraction of correctly classified cases

Classifier	Methods	Two groups (1–2A/2B–3)		Four groups (1/2A/2B/3)	
		Training (%) material I	Testing (%) material II	Training (%) material I	Testing (%) material II
Statistical multivariate	Object	83	62	51	19
	Texture	92	60	62	47
	Graph	78	78	57	47
Neuron networks	Object	97	55	63	40
	Texture	79	76	59	52
	Graph	81	83	61	57

Table 2
Reproducibility from the first to the second set of regions of interest in material II for the different classifier and feature extraction strategies

Classifier	Methods	Two groups (1–2A/2B–3)		Four groups (1/2A/2B/3)	
		Training (%)	Testing (%)	Training (%)	Testing (%)
Statistical multivariate	Object	78	76	67	52
	Texture	84	74	78	64
	Graph	85	77	68	58
Neuron networks	Object	70	64	70	33
	Texture	76	70	67	61
	Graph	90	77	75	65

classifier and a classification matrix for the neural networks. The two classifiers for each method were tested on the second group with 100 images from the same cases (test set). The result is shown in Table 2.

4.4. Combination of the three methods

In addition to studying and comparing the three approaches to image analysis based grading individually we tried to find a combination between the three that was better than any individual method. The aim was to be able to create a more powerful and robust classifier. For this purpose all significant features of the three methods from the training set in material II were combined to create two new classifiers, one based on statistical methods, the other on neural network methods. The stepwise linear discriminate analysis selected the following five features as the best ones for the four-group case: percentile coefficient of skewness of object area (Graph), difference average (Texture), difference variance (Texture), CPV_{8515} of standard deviation of major axis for neighbour objects (Object) and the standard deviation of the cluster area (Graph). Similarly the following features were automatically selected to be the best for the two-group case: PCS of object area (Object), peak transition probability (Texture), sum variance (Texture), the standard deviation of the cluster area (Graph) and CPV_{8515} of standard deviation of major axis for neighbouring objects (Object). As can be seen features from each of the three approaches are included in the top group for the neural network approach. It was not possible to study which features were the most important. However, the gain term of the back propagation neural network was -0.55 for the two-group and -1.55 for the four-group case. The resulting classifier accuracies are shown in Table 3.

Table 3
The classification accuracy and reproducibility for the combined features in material II

Classifier	Two groups (1–2A/2B–3)		Four groups (1/2A/2B/3)	
	Training (%)	Testing (%)	Training (%)	Testing (%)
Statistical multivariate	90	77	75	63
Neuron networks	94	80	71	67

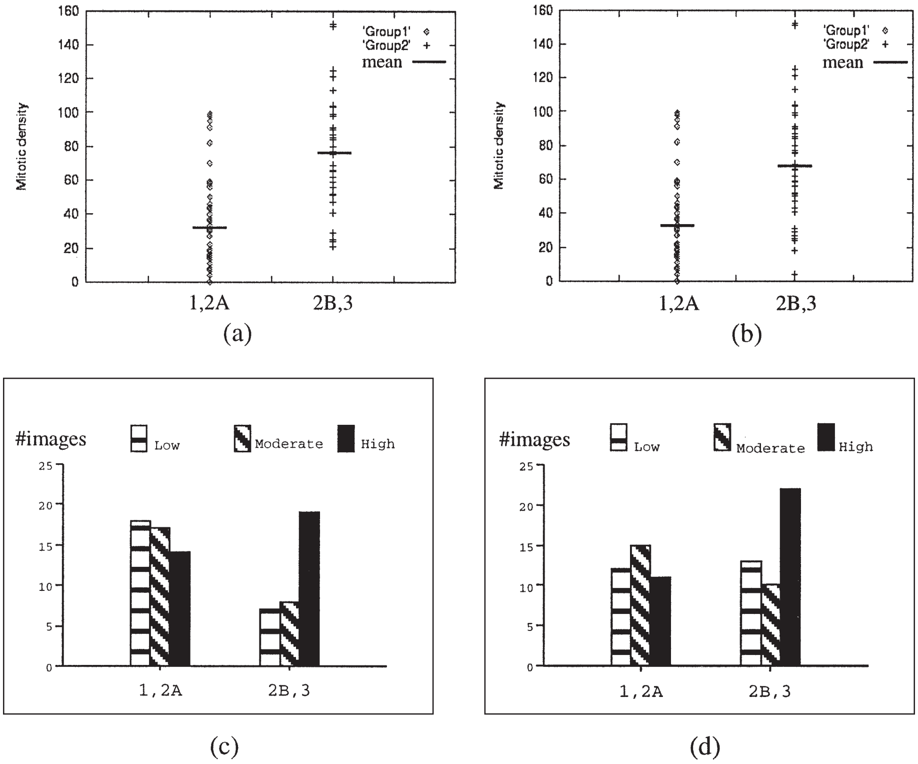


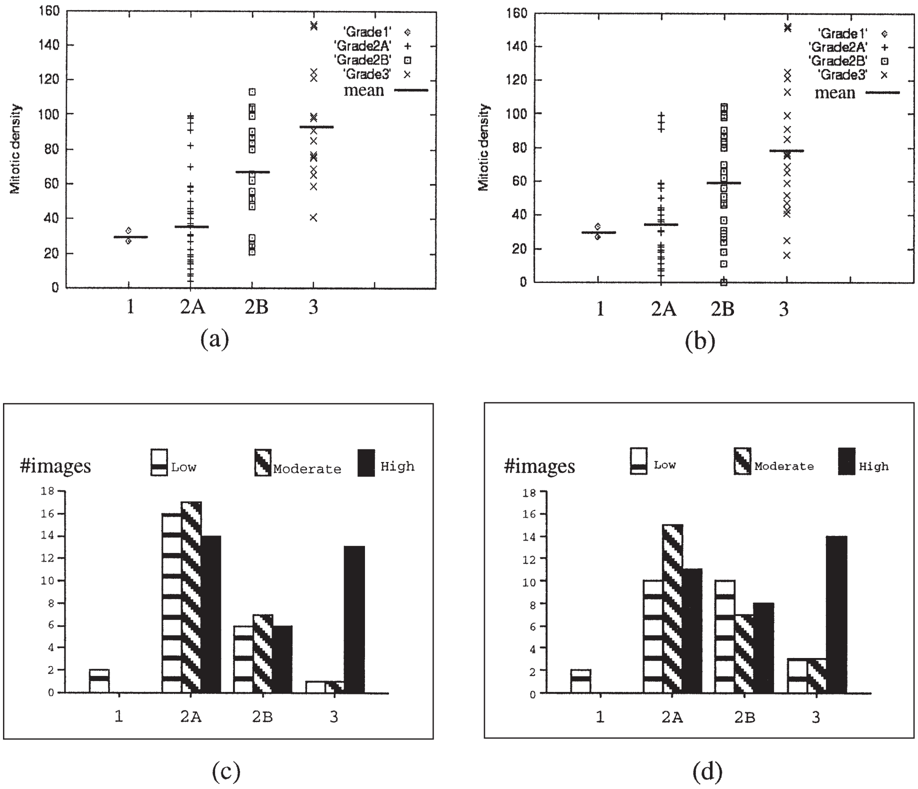
Fig. 8. Two groups: (a) Mitotic density related to subjective grading. Difference of mean is significant at $p < 0.01$. (b) Mitotic density related to computer grading. Difference of mean is significant ($p < 0.01$). (c) p53 expression classes (low, moderate, high) related to subjective grading. (d) p53 expression classes (low, moderate, high) related to computer grading.

4.5. Relation between subjective and image analysis based grading, mitotic density and p53 protein expression in material II

Analysis of the relation were made between subjective and the combined image analysis based grading for MD and expression of p53 protein in each specimen. The relation to MD and p53 expression for two groups are given in Fig. 8 and for four groups in Fig. 9.

5. Discussion

Quantitative image analysis of tumours started in the late sixties and is now a rapidly growing field. The correlation between nuclear size and tumour grade in bladder cancer was documented 1969 [27]



(a) (b)

(c) (d)

Fig. 9. Four groups: (a) Mitotic density related to subjective grading. Difference of mean is significant ($p < 0.01$) between Grade 2A and Grade 2B, Grade 2A and 3. (b) Mitotic density related to computer grading. Difference of mean is significant ($p < 0.01$) between Grade 2A and Grade 3. (c) p53 expression classes (low, moderate, high) related to subjective grading. (d) p53 expression classes (low, moderate, high) related to computer grading.

and confirmed in numerous investigations. Already in 1972 Prewitt [39] proposed a computerized grading scheme based on the analysis of a number of features extracted from digitized images of tissue sections. She later followed up on the idea in several other papers [40,41]. Van der Poel et al. [37] reviewed quantitative investigations of transitional cell bladder carcinoma.

The considerable increase of interest in this field in recent years undoubtedly has to do with the fact that computers and image digitizing equipment have now reached price/performance levels that make them feasible for use in some clinical routines. So there are good reasons to expect a significant increase in the use of computer assisted methods in pathology. Using computerized image analysis, many quantitative features that are not readily apparent to the human eye can be utilized, such as DNA content, high order texture descriptors, specific shape irregularities and geometrical structures together with the more visually obvious features. In this study we have developed and evaluated three different approaches to feature extraction and two different kinds of classification strategies. The feature extraction has been designed to capture both visually obvious as well as more abstract quantitative features. Each feature extraction approach has its advantages and problems.

The object oriented features are sensitive to a proper segmentation which is hard to do with high quality automatically. There are also difficulties with the influence of the 3D nature of the original tissue on the definition of what is an object in 2D. Another set of problems is caused by the varying staining density between different preparations. Through the use of dynamic selection of thresholds and postprocessing of the binary images we have minimized the influence of these problems. The influence of varying staining density is further reduced by the use of secondary features obtained from the initial feature distributions rather than the features as such. By using the distances between certain percentiles of the distributions the features are normalized against variations in absolute values.

Textures are in general very sensitive to focusing problems. Due to differences in the scale of the texture elements, this problem is more severe when the chromatin texture within the nucleus is to be analyzed than when the overall texture of the tissue section is to be analyzed as done in our study. Another difficulty in using textures is to find measures that are invariant to irrelevant variations in staining, image orientation, illumination level, etc., yet still provide reproducible differences between the different grades. To try to overcome these difficulties we used histogram equalization and orientation independent secondary features derived from the original oriented texture features. Still we did see some problems in reproducibility of the textural features between the different sets of specimens.

The third group of features, those obtained from the minimum spanning tree connecting the nuclear centroids in tissue sections share the sensitivity to segmentation problems with the object oriented features. Another problem relates to computational efficiency. Computing the MST is a computationally demanding task. However, several efficient implementations have been proposed for solving the minimum spanning tree problem [25,42]. We used Kruskal's algorithm with a time complexity of $\mathrm{O}(e \log e)$, where e is the number of edges in the graph. This lead to reasonable computation times. Nuclei lying at the image border is a problem when calculating the MST features. The border effect which is related to the number of nuclei in the regions is not evaluated in this study.

We also used two different classification strategies. Stepwise linear discriminant analysis and back propagation neural networks. An advantage of the linear discriminant method is that it indicates which features are most significant for the classifier. The method is however limited to linear classifiers. The network algorithm was susceptible to overlearning, as indicated by a peak in generalization performance followed by a consistent decline as the training continued. One disadvantage of using a neural network is that it is impossible to trace the significant features.

The combination of the three different kinds of features showed better classification results and higher reproducibility than most of the individual feature sets. It is also interesting to note that the automatic feature selection procedure in the stepwise linear discriminant analysis algorithm selected features from all three groups among the top five features. It thus seems that the different ways of extracting features provide complementary rather than fully redundant information.

The reproducibility of some grading systems can be as low as 30–40% [5,45]. Furthermore, in some tumours the correlation between clinical outcome and grading before treatment can be low or none [20,44]. However, grading of bladder carcinoma seems to have considerable prognostic potential despite the incongruity between the various systems used [5,30,33]. In our work we have shown that we can obtain fairly high and reproducible correlation between subjective and image analysis based grading, around 80% in the two-group case and around 65% in the four-group case. Almost as good reproducibility is achieved when the classifiers are trained on one material and system and tested on an entirely different material and system. The obtained computer grades also correlate well to mitotic density and p53 expression. However, we did not find any significant correlation between any of the features from the computer analysis and the mitotic density when we tested a direct multiple stepwise regression model, neither when the model was linear nor when it was logarithmic or squared. Similarly, neither the subjective nor the computer grading correlated time to recurrence. The follow up time was too short to evaluate time to progression and survival.

It is interesting to note that the features showing the best correlation to increment in grade were different measures of variation. This finding strengthens the notion that malignancy is a consequence of variability and disharmony of cellular organisation at all levels, and that increased malignancy grade represents an exaggeration of such disharmony.

In this study we have used surrogate endpoints for evaluation of prognostic power. We have mainly tried to mimic the subjective grade and studied correlations to MD and p53 expression. In preliminary studies on material I we did find an interesting relation between the texture based computer grade and patient survival time. This result will, however, need to be confirmed on independent material in the future as mortality data become available.

6. Conclusions

We have developed a computerized grading system for transitional cell bladder carcinomas. The combined method shows good correlation with the visual grading, mitotic density and p53 and demonstrates a robust prognostic power. The method could become a valuable addition to the tool available for estimating the prognosis of bladder carcinoma patients.

7. Acknowledgements

The authors gratefully acknowledge the support and constructive criticism from Dr B. Nordin and the other members of the Centre for Image Analysis staff and the Department of Pathology and Urology at the Uppsala University Hospital. Furthermore, we would like to thank the members of the Swedish–Norwegian Bladder Cancer Study Group for providing material II: P. Ekman, Karolinska Institute, Stockholm, C. Lundholm Västerås Central Hospital, J.O. Ohlsson, Karlstad Central Hospital, T. Lindeborg, Eskilstuna Hospital, H. Wijkström, Huddinge Hospital, Stockholm, S. Susskind, Hudiksvall, J. Sten, Mora, M. Lagerkvist, St Göran, Stockholm, S. Jansson, Örebro, K.J. Tveter, Ullevål Hospital, Oslo. This research was supported by the Swedish National Board for Technical Development under grant number 90-01178P and grant 2323-B87-01x from the Swedish Cancer Society and Lion’s Association, Uppsala.

References

[1] E. Bengtsson and O. Eriksson, The design of an image analysis system, in: *5th Scan. Conf. Image Analysis*, 1987, pp. 217–224.

[2] A. Bergkvist, A. Ljungqvist and G. Moberger, Classification of bladder tumors based on the cellular pattern, *Acta Chir. Scand.* **130** (1965), 371–378.

[3] G. Borgefors, Distance transforms in digital images, *Comput. Vision Image Process* **34** (1986), 344–371.

[4] A.C. Broders, Epithelioma of the genito-urinary organs, *Ann. Surg.* **75** (1922), 574–580.

[5] C. Busch, A. Engberg, B.J. Norlen and B. Stenkvist, Malignancy grading of epithelial bladder tumours, *Scand. J. Urol. Nephrol.* **11** (1977), 143–148.

[6] C. Busch and J. Vasko, Differential staining of mitoses in tissue sections and cultures cells by a modified methenamine-silver method, *Lab. Invest.* **59** (1988), 876–878.

[7] D. Cheriton and R.E. Tarjan, Finding Minimum Spanning Trees, *SIAM J. Comput.* **5** (1976), 724–742.

[8] H.-K. Choi, J. Vasko, E. Bengtsson, T. Jarkrans, P.-U. Malmström, K. Wester and C. Busch, Grading of transitional cell bladder carcinoma by texture analysis of histological sections, *Anal. Cell. Pathol.* **6** (1994), 327–343.

[9] H.-K. Choi, E. Bengtsson, T. Jarkrans, J. Vasko, K. Wester, P.-U. Malmström and C. Busch, Minimum spanning trees (MST) as a tool for describing tissue architecture when grading bladder carcinoma, in: *Proceedings of the 8th International Conference on Image Analysis and Processing* (*ICIAP ’95*), Sanremo, Italy, 1995, pp. 615–620.

[10] J.S. DaPonte and P. Sherman, Classification of ultrasonic image texture by statistical discriminant analysis and neural networks, *Comput. Med. Imaging and Graphics* **15** (1991), 3–9.

[11] W.J. Dixon, *BMDP Statistic Software Manual, 7M*, Department of Biomathematics, University of California Press, Berkeley, Los Angeles, Oxford, 1990.

[12] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

[13] C. Dussert, M. Rasigni, J. Palmari, G. Rasigni, A. Llebaria and A. Marty, Minimum spanning tree analysis of biological structures, *J. Theor. Biol.* **125** (1987), 317–323.

[14] J. Freeman and D. Skapura, *Neural Networks: Algorithms, Applications and Programming Techniques*, Addison-Wesley, 1992, pp. 89–125.

[15] G. Gomori, A new histochemical test for glycogen and mucin, *Am. J. Clin. Pathol.* **10** (1946), 177–179.

[16] M.S. Greenblatt, W.P. Bennet, M. Hollstein and C.C. Harris, Mutations in the p53 tumor suppressor gene: Clues to cancer etiology and molecular pathogenesis, *Cancer Res.* **54** (1994), 4855–4878.

[17] P.A. Hall and D.P. Lane, p53 in tumour pathology: can we trust immunohistochemistry?, *J. Pathol.* **172** (1994), 1–4 (revisited).

[18] R.M. Haralick, K. Shanmugam and I. Dinstein, Textural feature for image classification, *IEEE Trans. Syst. Man. Cybern.* SMC-**3** (1973), 610–621.

[19] T. Jarkrans, J. Vasko, E. Bengtsson, H.-K. Choi, P.-U. Malmström, K. Wester and C. Busch, Grading of transitional cell bladder carcinoma by image analysis of histological sections, *Anal. Cell. Pathol.* **8** (1995), 135–158.

[20] O. Johansson, J.E. Johansson, L.G. Lindberg and A. Sydsjö, Prognosis, recurrences and metastases correlated to histological cell type in carcinoma of the uterine cervix, *Acta Obstet. Gynec. Scand.* **55** (1976), 255–259.

[21] R.C. Johnson and D.W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice-Hall, 1992.

[22] K. Kayser, B. Kiefer and U. Burkhardt, Syntactic structure analysis of bronchus carcinomas – first results, *Acta Stereol.* **4** (1985), 249–253.

[23] K. Kayser, H. Stute, J. Bubenzer and J. Paul, Combined morphometrical and syntactic structure analysis as tools for histomorphological insight into human lung carcinoma growth, *Anal. Cell. Pathol.* **2** (1990), 167–178.

[24] K. Kayser, H. Stute and M. Tacke, Minimum spanning tree, integrated optical density and lymph node metastasis in bronchial carcinoma, *Anal. Cell. Pathol.* **5** (1993), 225–234.

[25] J.B. Kruskal, On the shortest spanning subtree of a graph and the traveling salesman problem, *Proc. Amer. Math. Soc.* **7** (1956), 48–50.

[26] P.A. Lachenbruch and M.R. Mickey, Estimation of error rates in discriminant analysis, *Technometrics* **10** (1968), 1–11.

[27] P.E. Levi, E.H. Cooper, D. Phil, C.K. Anderson, N. Path and R.E. Williams, Analyses of DNA-content, nuclear size and cell proliferation of transitional cell carcinoma in man, *Cancer* **23** (1969), 1074–1085.

[28] A.J. Levine, M.E. Perry, A. Chang, A. Silver, D. Dittmer, M. Wu and D. Welsh, The 1993 Walter Hubert lecture: the role of the p53 tumour-suppressor gene in tumorigenesis, *Br. J. Cancer* **69** (1994), 409–416.

[29] C. Lundholm, B.J. Norlén, P. Ekman, S. Jahnson, M. Lagerkvist, T. Lineborg, J.O. Olsson, K. Tveter, H. Wijkström, R. Westberg and P.-U. Malmström, A randomized prospective study comparing long-term intravesical instillations of mitomycin-c and bacillus calmette-guerin in patients with superficial bladder carcinoma, *J. Urol.* **156** (1996), 1–5.

[30] P.-U. Malmström, C. Busch and B.J. Norlen, Recurrence, Progression and survival in bladder cancer. A retrospective study of 232 patients with >5 years follow-up, *Scand. J. Urol. Nephrol.* **21** (1987), 185–195.

[31] R. Marcelpoil and Y. Usson, Methods for the study of cellular sociology: Voronoi diagrams and parametrization of the spatial relationships, *J. Theor. Biol.* **154** (1992), 359–369.

[32] B. Nordin, The development of an automatic prescreener for the early detection of cervical cancer: Algorithms and implementation, Ph.D. Thesis, University of Uppsala, 1989.

[33] E.C.M. Ooms, W.A.D. Anderson, C.L. Alons, M.E. Boon and R.W. Veldhuizen, An analysis of the performance of pathologists in the grading of bladder tumors, *Hum. Pathol.* **14** (1983), 140–143.

[34] E.C.M. Ooms, P.H.J. Kurver, R.W. Veldhuizen and C.L. Alons, Morphometric grading of bladder tumors in comparison with histologic grading by pathologists, *Hum. Pathol.* **14** (1983), 144–150.

[35] C.L. Orbert, E.W. Bengtsson and B.G. Nordin, Watershed segmentation of binary images using distance transformations, in: *Proc. Image Process: Nonlinear Image Process IV*, SPIE 1902, 1993, pp. 159–170.

[36] R.P.E. Pauwels, R.F.M. Schapers, A.W.G.B. Smeets, F.M.J. Debruyne and J.P.M. Geraedts, Grading in superficial bladder cancer: (1) Morphological criteria, *Br. J. Urol.* **61** (1988), 129–134.

[37] H.G. van der Poel, H.E. Schaafsma, G.P. Vooijs, F.M.J. Debruyne and J.A. Schalken, Quantitative light microscopy in urological oncology, *J. Urol.* **148** (1992), 1–13.

[38] N.J. Pressman, Markovian analysis of cervical cell images, *J. Histochem. Cytochem.* **24** (1976), 138–144.

[39] J.M.S. Prewitt, Objective characterization of tissue sections by digital image processing, in: *Proc. 17th Ann. Conf. Eng. Med. Biol.*, 1972, p. 481.

[40] J.M.S. Prewitt, On some applications of pattern recognition and image processing to cytology, cytogenetics and histology, Ph.D. Thesis, University of Uppsala, 1978.

[41] J.M.S. Prewitt, Canonical representations for tissues and textures, *IEEE Computer Society* (1979), 470–476.

[42] R.C. Prim, Shortest connection networks and some generalizations, *Bell System Technical J.* **36** (1957), 1389–1401.

[43] H. Schimmelpenning, U.G. Falkmer, K. Hamper, G. Seifert and G.U. Auer, Variations in Feulgen stainability of epithelial parenchymal cells extracted from paraffin-embedded salivary gland specimens, *Cytometry* **11** (1990), 475–480.

[44] U. Stendahl, H. Willen and R. Willen, Classification and grading of invasive squamous cell carcinoma of the uterine cervix, *Acta Radiol. Oncol.* **18** (1979), 481–495.

[45] U. Stendahl, H. Willen and R. Willen, Invasive squamous cell carcinoma of the uterine cervix II: reproducibility of a histopathological malignancy system, *Acta Radiol. Oncol.* **20** (1981), 71–76.

[46] J. Vasko, P.-U. Malmström, A. Taube, K. Wester and C. Busch, Towards an objective method of mitosis counting and its prognostic significance in bladder cancer, *J. Urol. Pathol.* **3** (1995), 315–326.

[47] L. Vincent and P. Soille, Watersheds in digital space: an efficient algorithm based on immersion simulations, *IEEE Trans. Pattern Analysis and Machine Intelligence* **13** (1991), 583–598.

[48] Q. Zhan, I. Bae, M.B. Kastan and A.J. Fornace, The p53-dependent gamma-ray response of GADD45, *Cancer Res.* **54** (1994), 2755–2760.