Cytometry 12:602–613 (1991)

Towards a Quantitative Grading of Bladder Tumors[1]

Uneco de Meester, Ian T. Young[2], Jan Lindeman, and Hans C. van der Linden

Pattern Recognition Group, Faculty of Applied Physics, Delft University of Technology, Delft, The Netherlands (U.d.M, I.T.Y.); Institute of Pathology of Slotervaart Hospital, Amsterdam, The Netherlands (J.L.); and Department of Pathology of the Free University Hospital, Amsterdam, The Netherlands (H.C.v.d.L.)

Received for publication January 25, 1991; accepted May 30, 1991

The histological inspection of tumor tissue for the purpose of reporting a tumor grade is a problem of significant clinical importance. The grading by a pathologist is only partly reproducible due to vaguely defined, subjective criteria. In this article we describe and evaluate a set of measurable features that quantitate the differences in tumor tissue. Different aspects of the reproducibility of the measurements under varying conditions of image selection, focus, and noise have been investigated.

Three hundred thirty-three images were digitized from 111 bladder tissue sections (4 μm thick, Feulgen stained), using the ICAS microscope-camera platform. A segmentation routine was developed to segment the images into nuclei and background without any user interaction. Size, shape, optical density, and texture features were measured on and among the *objects* found by this segmentation routine using the image analysis package Acuity.

The results of the measurements showed that there is a significant quantitative difference between grade 1 and grade 3 tumors. Grade 2 tumors can be described as "in between grade 1 and grade 3" and falling somewhere on an increasing scale between grades 1 and 3. Grade 2 tumors do not seem to represent a statistically distinct population.

The procedure described here is shown to be quite reproducible in the presence of noise, reasonably reproducible in the event of a modest amount of defocusing (with grade 3 tumors exhibiting the most sensitivity), and less reproducible in the context of *which* fields-of-view are chosen for analysis.

Key terms: Tumor grading, epithelial bladder tumors, tissue architecture, image segmentation, image analysis, histometry, quantitative microscopy

The aim of this study is to develop a quantitative description of bladder tumor tissue that could be used as an objective and reproducible grading system in an automated or semi-automated system. For tumor grading, both the architecture of the tumor tissue as well as cell features such as size, shape, texture, and optical density of the cell nuclei are taken into account. The definitions that are traditionally used for the various grades are vague and admit subjective interpretation. Therefore the inter- and intra-observer reproducibility is low (4, 20).

At the cellular (cytologic) level, the cellular objects are described in terms of geometric (size), morphological (shape), and densitometric properties (assuming an absorptive stain). In higher grade tumors, cell nuclei are more irregular in terms of size, shape, texture, and optical density. Further, the grade of malignancy seems to correlate with nuclear enlargement (10, 15). Accurate measurements of nuclear size and shape are relatively easy to ascertain and can supply significant objective information (3).

At the tissue (histological) level, biological objects are described in terms of their architecture. The spatial organization of the cells tends to be more irregular in

[1]This work was partially supported by The Netherlands Foundation for Medical Research, Medigon (grant 900-538-016) and by the SPIN research program, Three-Dimensional Image Analysis.

[2]Address reprint requests to Professor Ian T. Young, Pattern Recognition Group, Faculty of Applied Physics, Delft University of Technology, Lorentzweg 1, 2628 CJ Delft, The Netherlands.

the higher grade tumors. Various studies have described methods to *measure* this irregularity (5, 17, 28). Most of these techniques start with a segmentation algorithm to find the nuclei in the background and then divide the image into compartments using a skeleton or other tessellation algorithm (1). For the compartments found in this way, a number of size, shape, and distance properties can be computed. The variation of these properties over the image can be used as a measure of differentiation between regular and irregular tumor architectures.

Another approach is to use the distribution of the inter-object distances to model the differences in tissue architecture. The use of the inter-object distances has great advantage over measurements on the image regions found via the background skeleton (22). Where image regions change considerably when objects are missed or when the image contains an artefact such as a Brunn's nest, the nearest neighbor distances of the objects remain almost the same.

Combined Approach

Both cellular level and architectural level measurements suffer from scene segmentation problems. Intensive user interaction could be a solution but this would be in direct conflict with the primary goal of this project, which is the development of an objective system for the description and grading of tumors. The information available in a tissue section is a combination of effects at the cellular level *and* the architectural level. It may, therefore, be inappropriate to try to measure these effects separately. The problems encountered in a segmentation phase are *caused* by the phenomena that must be quantified.

Initially we intended to indicate all doubtful objects (e.g., conglomerations of two or more cells) manually and to exclude these objects from further measurements. Gradually it became apparent that these objects did contain the information to be measured. The pathologist associates overlapping and touching nuclei with an irregular tissue architecture. A review of the cytological criteria for malignancy supported this idea (10). Therefore, measurements of size, shape, intensity, and intensity distribution (texture) of objects—not necessarily single cell nuclei—were used to quantify the combination of architectural and cell level differences.

MATERIALS

Bladder biopsies from 1987 and 1988 were collected from the archives of the S.S.D.Z. (Stichting Samenwerkende Delftse Ziekenhuizen). From this material 4 μm thick tissue sections were made and stained according to the Feulgen pararosaniline procedure. All sections were inspected by an expert for a second opinion in addition to the routine diagnosis that was available from the archives. The routine diagnoses (in 1987 and 1988) were based upon hematoxylin eosin stained samples while the second opinions (in 1989) were based primarily upon the Feulgen-stained samples with incidental use of the original hematoxylin eosin preparations. The second opinion differed from the first opinion by one grade in 42 out of the 111 sections or about 38% of all sections. Neither grade 1 to 3 nor grade 3 to 1 differences of opinion occurred. Table 1 gives an overview of this learning set; the results are in agreement with results found in the literature (4). The material that is used in the rest of this manuscript was assigned the grades listed in Table 1. The samples lying along the main diagonal were called 1, 2, and 3; the naming of the off-diagonal samples will be discussed later. On all slides the area that was considered representative for the tumor for this study was indicated by the (second) expert. From each of the 111 sections used in this study, 3 images were collected for a total of 333 images.

Table 1
Confusion Matrix Demonstrating the Variability in Subjective Grading[a]

	Observer 2			
	Grade 1	Grade 2	Grade 3	Total
Observer 1				
Grade 1	**17**	6	0	23
Grade 2	22	**35**	7	64
Grade 3	0	7	**17**	24
Total	39	48	24	111

[a]Agreements between the two expert observers are indicated in boldface.

Hardware Instrumentation: Microscope and Camera

A Nikon Diaphot TMD inverted microscope was used with a Zeiss 63× neofluor dry objective (NA = 0.9) and a 1× TV relay lens as "eyepiece." The specimen was illuminated with "white" light from a 100 W quartz-halogen bulb filtered with a band pass color filter, whose maximum transmission wavelength was approximately 545 nm and whose bandwidth (full width, half maximum) was 60 nm (Nikon GIF). The condenser aperture was NA = 0.6 and adjusted for Köhler illumination. An infra-red blocking filter was placed in the light path in front of the CCD camera to prevent IR light from degrading the contrast.

The images were scanned using a Photometrics 200 Peltier-cooled CCD camera with 576 × 384 pixels and 12 bits per pixel (Photometrics Inc., Tucson, AZ). The noise performance of the camera is limited only by the photon statistics in each CCD well. This leads to a signal-to-noise (SNR) ratio in excess of 58 dB. The sampling density was measured at 0.351±0.00068 μm/pixel in both directions. The measurements in this research concentrate on cell nuclei of approximately 8 μm in diameter (≈ 23 pixels). Given an objective lens with a specific numerical aperture (NA = 0.9) and an illumination wavelength (λ = 0.550 μm), theory (26) says that a sampling distance d_s should be used that is λ/(4·NA) = 0.153 μm/pixel. The images have been (in-

tentionally) undersampled and the effect of this will be discussed later.

The CCD camera is (via camera electronics) controlled by a 68010-based dedicated microcomputer on which some image processing is possible. This system lacks sufficient flexibility, however, to serve as a real workbench for image processing and image analysis. Image acquisition and shading correction are the only functions performed using the functions of the CC-200 system. The images are then transferred via an IEEE-488 (GP-IB) interface to a SUN 3/110 workstation for further processing. A schematic overview of the system is shown in Figure 1.

Software Instrumentation: Workbench for Image Processing/Analysis

The general workbench for imaging processing used in this research was the TCL-Image processing package (Multihouse TSI, Amsterdam, The Netherlands) coupled to the Acuity image analysis package (27). The former was used for image segmentation and the latter for image measurement (feature extraction). Sample images of grade 1, 2, and 3 tumors as displayed under TCL-Image are shown in Figure 2.

METHODS

Shading Correction

Shading correction consisted of recording two test images through the camera $\mathbf{I}_w(x,y)$ and $\mathbf{I}_b(x,y)$. The former was a clear "white" field (that is, no specimen in the light path) and the latter a clear "black" field (that is, with no illumination). For each subsequent image, $\mathbf{I}_r(x,y)$, that was recorded, the following algorithm was applied to produce the corrected image $\mathbf{I}_c(x,y)$.

$$\mathbf{I}_c(x,y) = \frac{\mathbf{I}_r(x,y) - \mathbf{I}_b(x,y)}{\mathbf{I}_w(x,y) - \mathbf{I}_b(x,y)} \mathbf{I}_{avg}, \quad (1)$$

where $\mathbf{I}_{avg}$ was the intended average intensity of the "white" image.

Segmentation

The segmentation of tumor tissue images into the constituent cell nuclei is difficult for a number of technical as well as biological reasons:

1. A tissue section is a two-dimensional representation of a three-dimensional phenomenon. When tissue sections are used, partial and overlapping nuclei are unavoidable.
2. Because the section has a thickness of 4 μm, objects are not confined to the same focal plane. In practice a pathologist uses this fact to examine overlapping nuclei but for an automatic segmentation routine it can introduce considerable problems. The depth of focus of the optical system used in this research was approximately 3 μm [defined by Piller's formula (16)].
3. The staining of a tissue section is never perfectly homogeneous.

Further, in the context of this application, it is impossible to define a good evaluation criterion because it is impossible to define what is a *perfect* segmentation mask for an image of a tissue section. In addition, all tissue images are different and the technique that solves problems in one image may cause problems in another. The routine that was chosen gave acceptable results (evaluated visually) and is intuitively attractive. The goal of this study was to find a quantitative description of bladder tumors. It was not at all certain that a perfect segmentation mask was necessary as long as it was "equally imperfect" for different images.

The only knowledge one really has of the cell nuclei is that they appear as dark objects on a lighter background. Using this knowledge our approach to the segmentation of the tissue images involved two major steps: a grey-value processing of the image followed by a binary (morphological) processing.

Grey-Value Processing

For the reasons mentioned above, the contrast (the grey value of object pixels in relation to the grey value of background pixels) varies considerably over the image. Therefore global thresholding will not yield satisfactory results. When the *local* contrast is sufficient, however, a dynamic threshold (i.e., a threshold that varies over the image) can be the solution to this problem. The general strategy is to divide the entire image into a series of adjacent, non-overlapping windows. Within each window a threshold is determined and bilinear interpolation is then used to produce an image of the threshold as a function of position for the entire (576 × 384) image. This technique was first introduced by Chow and Kaneko in 1972 (6). The technique that we used to produce the threshold per window was the Isodata algorithm of Ridler and Calvard (19).

The size of the windows is an important parameter in this method. Choosing a value for this parameter always means a trade-off between preserving detail and robustness. In this work, a window size of 32 × 32 pixels was used; a size about 50% larger than the average nucleus.

Binary Processing

Binary image processing was used to further improve the resulting mask image. Readers not familiar with the operations used in binary (morphological) processing are referred to Serra (21), Giardina and Dougherty (13), or Haralick (14). The binary image processing steps that were used were:

1) Filling. The object masks, determined from the dynamic thresholding, are filled by propagating a white border into the *inverted* mask-image. The inverse of the result is an image that contains object masks without holes.

2) Skeletons. After the filling operation two skeletons of the *background* are computed (with and without the preserve-endpixel-condition). These skeletons are used to restore the topology of the background after the closing operations in the next step.

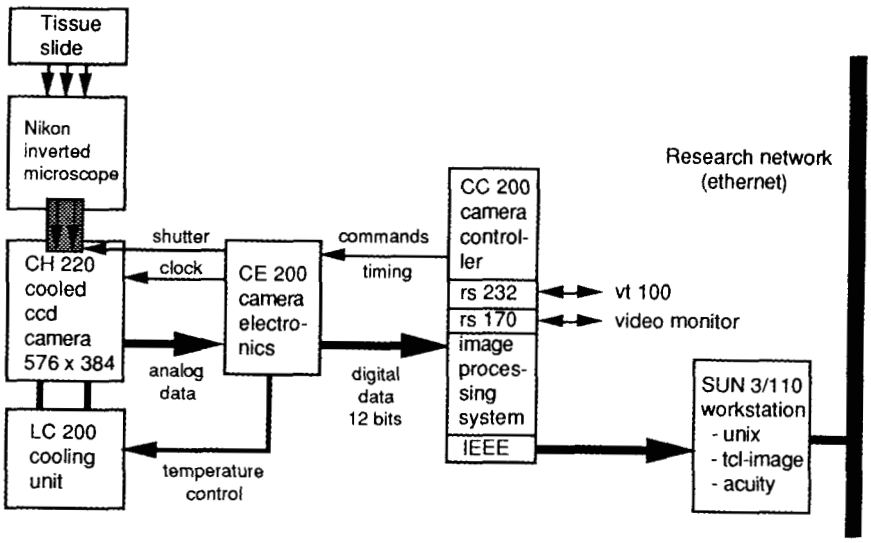


FIG. 1. Schematic overview of the ICAS system (24) used for image acquisition, processing, and analysis.

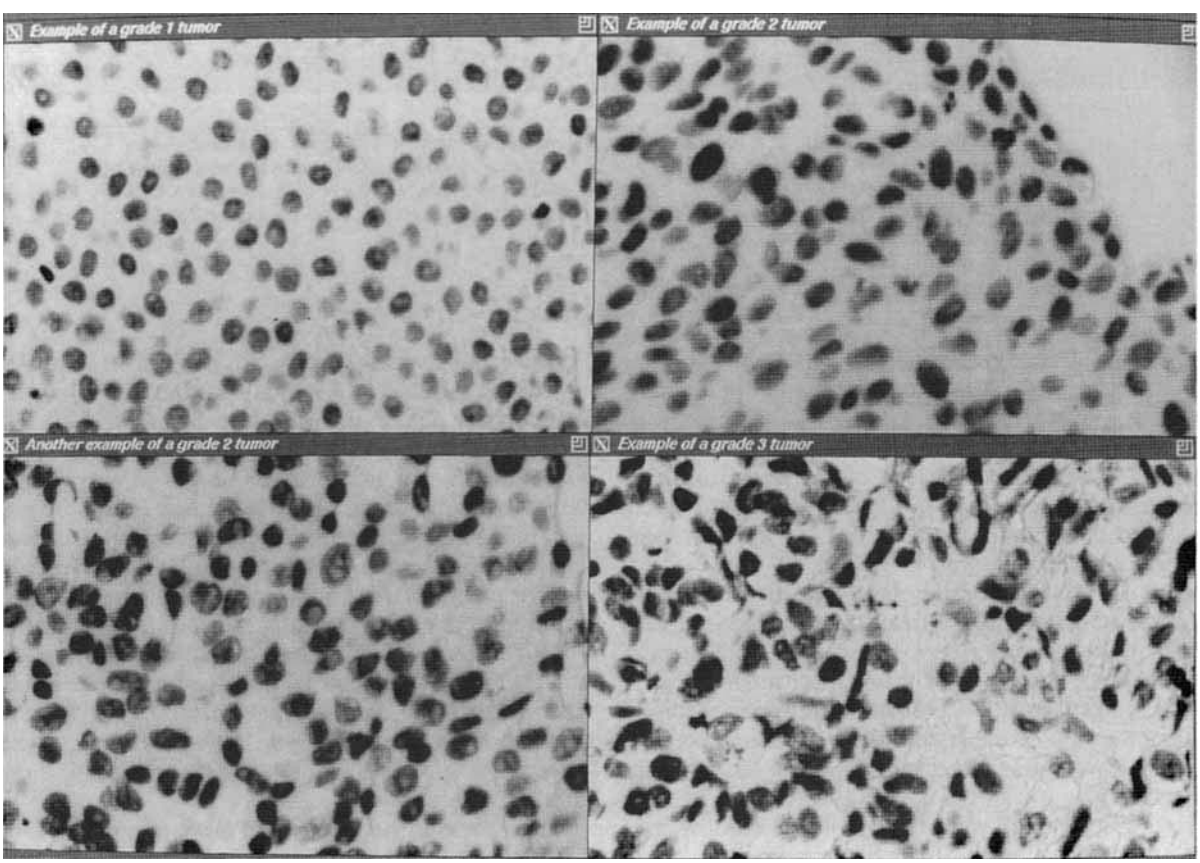


FIG. 2. Both observers agreed on the classifications for each of the following bladder tumor images. **Upper left,** example of a grade 1 image; **upper right,** example of a grade 2 image; **lower left,** example of a grade 2 image; and **lower right,** example of a grade 3 image.

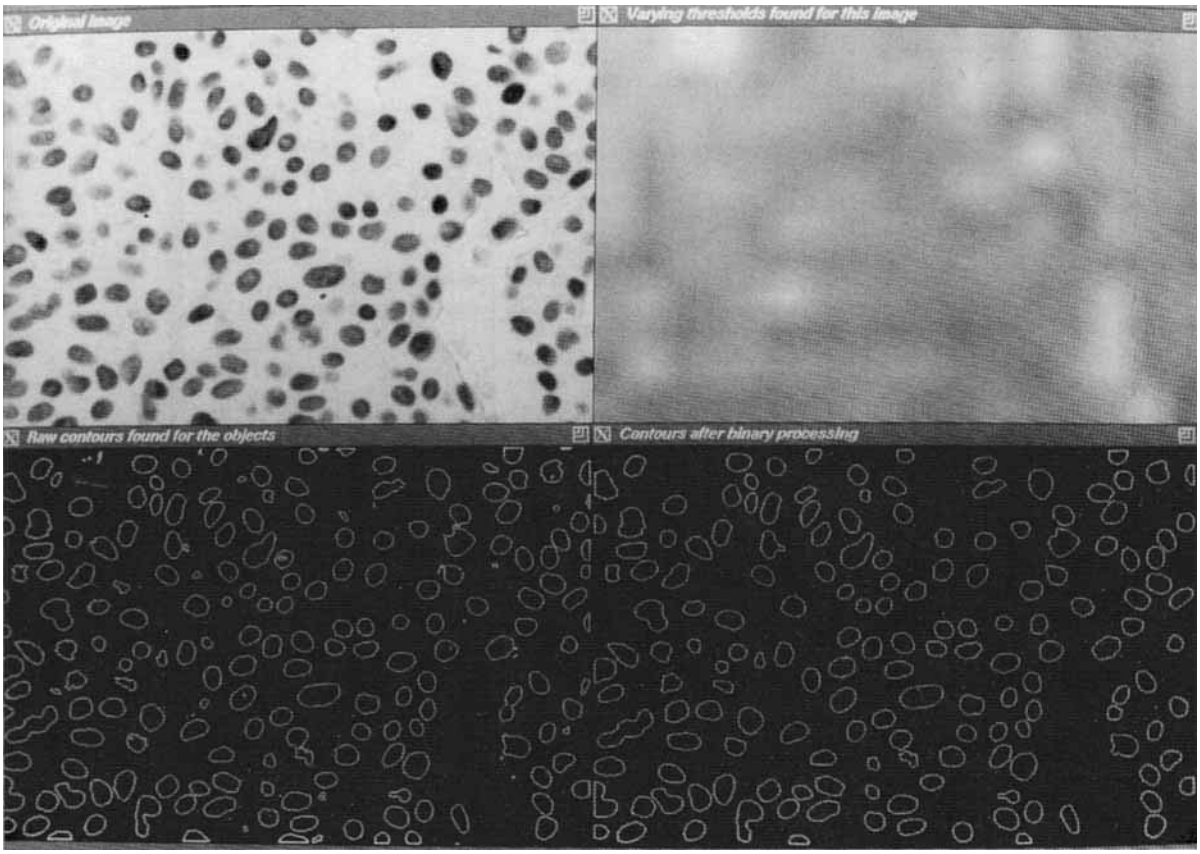


FIG. 3. **Upper left,** example of a bladder tumor image; **upper right,** the dynamic threshold determined by a combination of the Isodata algorithm in 32 × 32 windows and the subsequent bilinear interpolation; **lower left,** the resulting binary contours; and **lower right,** the binary contours as a result of steps 1–6. Note that there is a distorted object contour in the bottom right corner of the image. For this reason, the final step, 7, was necessary.

3) Closing. Concavities in the object masks are closed with an octagonally shaped structuring element of size 5 × 5.

4) Restore topology. By means of the skeletons computed in step 2 the topology of the background is restored using an *exclusive or*. This means that objects that became connected as a result of the closing operations are separated again.

After these operations the object masks do not contain holes and most of the concavities, caused by residual thresholding errors, have been repaired. The next steps aim to cut apart slightly touching nuclei.

5) New background skeleton. After one erosion of the image using a structuring element based upon the four nearest neighbors, the *majority-vote* filter is applied twice. The skeleton of the background is then computed. Some small objects do not survive the majority voting. The objects that have survived are propagated into the object masks that resulted from steps 1–4. The new background skeleton is now used to cut apart the loosely connected (i.e., touching) objects. (Note: *majority-vote* filtering in a binary image sets the center pixel in a 3 × 3 window to the value determined by the majority of pixels in that window.)

6) Removing small objects. The operation of step 5 also cuts off small extensions of the objects. These extensions are removed from the mask image by an erosion using an octagonally shaped structuring element of size 5 × 5 followed by a *pepper and salt* removal. (Note: *pepper and salt* filtering in a binary image removes isolated "1"s and isolated "0"s in a 3 × 3 window.)

7) Remove objects near the border. The shape of the final objects close to the border of the image is distorted due to the interaction of the previously described structuring elements with the image boundary. Objects within four pixels of the image boundary are therefore removed.

The effects of the grey value processing as well as the binary processing are illustrated in Figure 3a–d.

Although some vague (low contrast) nuclei are missed it is clear that the method adapts the threshold to the local contrast.

Features Measured

Tumor grading starts with *one* patient and ends with *one* judgement of the tumor grade for each tumor. Between the start and the end a substantial amount of information is taken into consideration in order to come to the final decision. The information starts with 18 features that the Acuity program measures directly or derives from each object in the image. These features, whose definition can be found in Young and Roos (27), can be separated into five categories: **SIZE**– area, perimeter, length and width; **SHAPE**– p2a, bending energy, normalized mean absolute curvature, sphericity, and eccentricity; **BRIGHTNESS**– optical density (OD) and OD per unit area (MOD); **TEXTURE**– heterogeneity, condensation, clumpiness, and margination; and **GENERAL**– center-of-mass, user classification, class, and angle.

The results of the measurements on ±150 objects per tissue are compressed into the statistics (mean and $CV = \sigma/\mu$) of 13 distinct features measured on each tissue section; see Figure 4. (The features length, width, angle, and center-of-mass are not directly used.) It is important to emphasize that the results of the measurements on objects in *three* images are all measurements from the same section.

Tissue architecture is described by the distribution of nearest neighbor distances (7, 23). While the use of measures based upon the tesselation of a tissue section through the use of the exo-skeleton or Voronoi diagram is quite sensitive to preparation artefacts or irrelevant anatomic features, the distance from a cell nucleus in a tissue section to the nearest other cell nucleus is much more stable. These distances were calculated by a separate C program from the positions of the center-of-mass of each object as measured by Acuity.

Tools for Data Analysis and Data Handling

The experiments yielded one feature vector for each sample that was processed. In this research the mean value of a certain feature and its CV are regarded as separate features for a feature selection procedure. For example mean area expresses the mean size of the cell nuclei within a population while the CV of area is a measure of the relative variation in size within that population.

To analyse this data, software for data manipulation, multivariate statistical analysis, and pattern recognition was needed. The algebraic manipulation of data acquired from Acuity was performed under Excel 2.2 (Microsoft Corporation, Redmond, WA). The software package ISPAHAN (Department of Medical Informatics, Erasmus University, Rotterdam) was used for multivariate statistical analysis and pattern recognition (12). The package is capable of handling feature spaces up to a dimension of 45 and offers a variety of statistical methods and classification techniques via a menu driven user interface.

For the handling of the multidimensional data a software package, MacSpin (D^2 Software Inc., Austin, TX) was used. This package has facilities for data display, data inspection, and data manipulation through a pseudo-3D visual presentation of various combinations of features and data (sub)sets. Transformations and projections (i.e., linear combinations) can be easily computed and inspected.

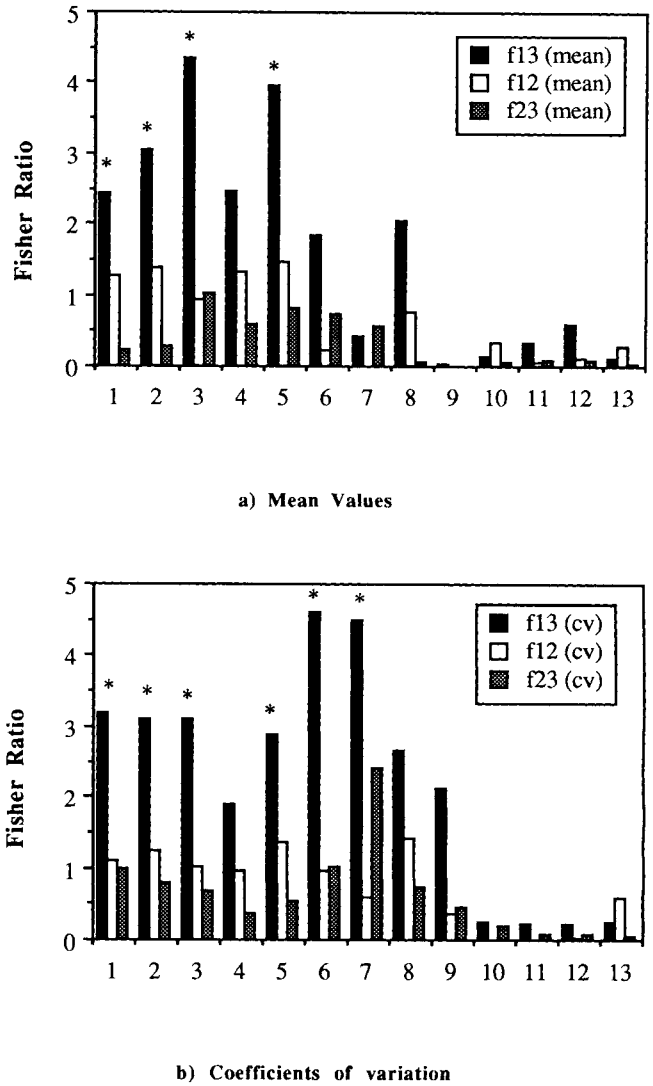


FIG. 4. Fisher ratios for the various features. In (**a**) the mean values and in (**b**) the coefficients-of-variation. The features are numbered 1, area; 2, perimeter; 3, p2a; 4, bending energy; 5, normalized mean absolute curvature (nmac); 6, sphericity; 7, eccentricity; 8, optical density (stain); 9, mean optical density (stain per area); 10, heterogeneity; 11, condensity; 12, clumpiness; and 13, margination. The ten highest values are marked by the asterisks (*).

RESULTS

It can be inferred from the descriptions pathologists use for the different tumor grades (10) that grade 2 lies between grade 1 and grade 3 *in all respects*. Therefore a combination of features that discriminates well be-

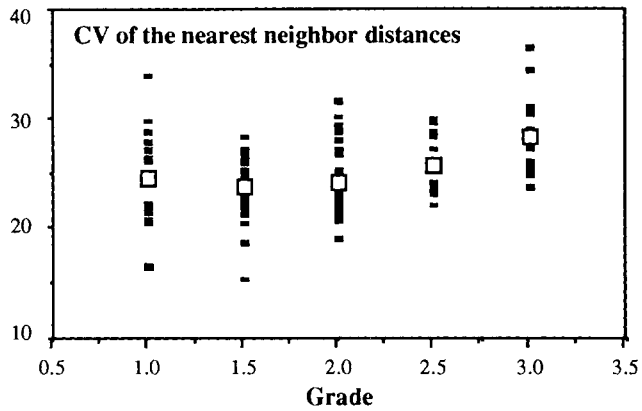


FIG. 5. The distribution of the CV of the nearest neighbor distances for each of the 111 tumors in Table 1. The means for each grade are indicated by □.

tween grade 1 and grade 3 tumors may also be used to define a continuous scale *from* grade 1 *to* grade 3 onto which the grade 2 tumors can be mapped. To find this discriminating combination of features only the measurements on the clearly distinct grade 1 and grade 3 tumors were used. The tools that we have described above have been used to determine this continuous scale.

Univariate Analysis of Features

The Fisher ratio (11) was used to evaluate the discriminating power of the various individual features. The Fisher ratio f_{xy} between grades x and y is defined as

$$f_{xy} = \frac{(m_x - m_y)^2}{(s_x^2 + s_y^2)}. \quad (2)$$

For the tumors of grade x, m_x is the sample mean, s^2_x the sample variance of a given feature (for example, mean perimeter, CV of area), and these sample statistics are calculated with N_x samples. For example, f_{12} is calculated using $N_1 = 17$ grade 1 tumors and $N_2 = 35$ grade 2 tumors. The results of this univariate analysis are shown in Figure 4.

Figure 5 shows the distribution of the CV of the nearest neighbor distance for each of the tumor grades. A number of preliminary conclusions can be drawn from Figures 4 and 5.

1. The difference between grade 1 tumors and grade 3 tumors (f_{13}) is considerable for certain size and shape features. The grade 2 tumors overlap the grade 1 and grade 3 tumors at Fisher ratio ≤ 2. That is, f_{12} and f_{23} are, in most cases, less than two when only one feature is taken into account. The value two is an ad hoc choice based upon the distribution displayed in Figure 4.
2. Neither the mean values nor the CVs of the texture features discriminates between the classes.
3. The mean stain and mean area provide some discriminating power between grade 1 and grade 3; the mean stain per area does not.
4. Both mean values and CVs of the features calculated by Acuity are relevant. For one feature (e.g., perimeter), the mean value provides the most discriminating power and for another (e.g., sphericity), the CV (indicating a difference in terms of homogeneity-heterogeneity) provides this power.
5. It is clear from Figure 5 that the CV of nearest neighbor distances, a possible measure of tissue architecture, is not capable of providing tumor grade discrimination in this experiment.

Table 2
The Set of Four Features as Selected by the Feature Selection Procedure AUTMSK of the Software Package ISPAHAN[a]

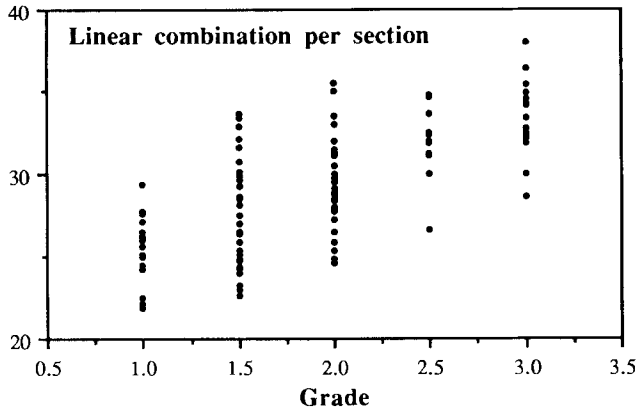
Feature	F-value	Weight
CV of sphericity	78.07	0.731
Mean perimeter	3.71	0.296
CV of P2A	2.41	−0.567
CV of area	1.87	0.238

[a]The weight for each feature was calculated by a Fisher projection. The weights apply to normalized versions of each feature where the mean of the ith feature is zero ($\mu_i = 0$) and the standard deviation is one ($\sigma_i = 1$).

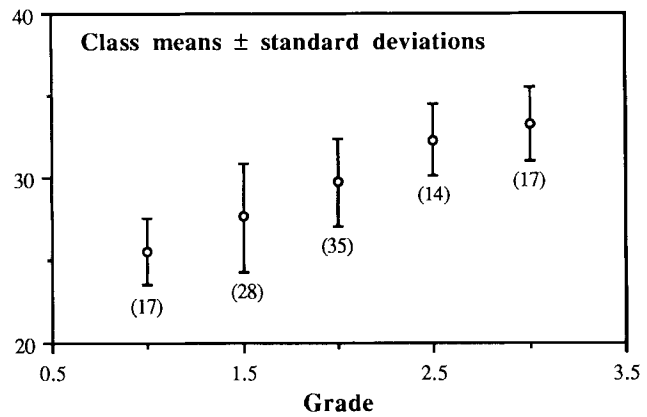
Multivariate Feature Selection

From the 26 features measured, the ten best discriminating ones were chosen on the basis of their univariate behavior (see Figure 4.) We propose to use a linear combination of features to achieve a discrimination between the various tumor grades. The ten chosen features were evaluated by the ISPAHAN procedure AUTMSK (12) in order to choose the "best" combination of four. Again the grade 1 ($N_1 = 17$) and grade 3 ($N_3 = 17$) tumor data were used. The procedure is essentially stepwise linear discriminant analysis where a feature is added to the set of those already chosen based upon its F-value (8). (Note that the F-value is *not* the same as the Fisher ratio.) The set that was selected consisted of two CV-based shape features, one CV-based size feature, and one mean-based size feature. The F-values as well as the weights assigned to each of the four (normalized) features are given in Table 2.

Figure 6 presents graphs of the values of this linear combination versus the grade. The 28 samples that were graded as "grade 1" by one observer and "grade 2" by the other have been given the grade "1.5". The 14 samples that were graded as "grade 2" by one observer and "grade 3" by the other have been given the grade "2.5". Although the classes overlap it is clear that the mean values (of the linear combination of features for a given class) show a monotonic increasing trend from grade 1 towards grade 3. This graph supports the idea that it is possible to define a quantitative, continuous scale between the grade 1 and grade 3 tumors. It must



(a) Linear combination of four features per section



(b) Mean and standard deviation per grade

FIG. 6. The values of the linear combination of Mean_perimeter, CV_spher, CV_p2a, and CV_Area for each of the 111 samples in Table 1. (a) The value of the linear combination for each of the 111 samples. (b) The mean and standard deviation for each of the five grades; the number of samples for that grade is given in parentheses. The grades are explained in the text. Although the classes overlap it is clear that the class means show a monotonic trend from grade 1 towards grade 3.

be kept in mind that this quantitative scale is a linear combination of only the four features mentioned above.

The differences between the means of the five grades 1, 1.5, 2, 2.5, and 3 show statistical significance at varying levels based upon use of the Student's t test between means with unknown and unequal variances (18). The results are given in Table 3. From the results in Table 3 we see that there is at least a factor of ten more significance associated with the "distinct" populations 1, 2, and 3 than with the "hybrid" populations 1.5 and 2.5. These results support the hypothesis that the ordering observed in Figure 6b does not occur by chance.

Table 3
Student's t Test Comparisons of the Various Populations Means Illustrated in Figure 6b[a]

Populations	t value	DOF	Significance
1 vs. 2	6.19	43	$\alpha < .001$
2 vs. 3	4.95	40	$\alpha < .001$
1 vs. 3	10.47	34	$\alpha < .001$
1 vs. 1.5	2.64	45	$\alpha < .025$
1.5 vs. 2	2.69	54	$\alpha < .01$
2 vs. 2.5	3.45	31	$\alpha < .01$
2.5 vs. 3	1.17	30	$\alpha < .20$

[a]The varying degrees-of-freedom (DOF) are associated with this particular version of the t test.

DISCUSSION

Sampling Density

The images used in this study were sampled at a distance (0.351 μm/pixel) that is incompatible with the Nyquist rate, a sampling distance that would correspond to 0.153 μm/pixel. To have satisfied the conditions of the Nyquist theorem (26), a sampling density approximately three times higher than the one used for this study would have been needed. This in turn would have meant that each image would have required nine times as many pixels. The alternative would have been to study a region that encompassed one ninth of the area in one of our current images. This would have meant far fewer cells per image and less statistical certainty.

The increase in image fidelity afforded by the increased sampling density would undoubtedly have provided more accurate measurements (25). Use of the Nyquist sampling density does not, in and of itself, guarantee completely accurate measurements. That is, measurements made below the Nyquist frequency are of a certain accuracy, those made at the Nyquist frequency of a higher accuracy, and those made at a still higher frequency, a still higher measurement accuracy. The issue thus becomes: if the measurement accuracy at a given sampling density is sufficient to the task then the issue of the Nyquist frequency is not pivotal.

What has been sacrificed through the choice of sampling density was the ability to use features that related to the structure of chromatin distribution within a cell nucleus. The sampling was simply too coarse for that (29).

Data Analysis

ISPAHAN and MacSpin proved to be a powerful combination for the evaluation of multidimensional data. We started with a set of 26 features (13 means and 13 CVs), reduced them to a set of ten with the largest Fisher ratio, and then used stepwise linear discriminant analysis to choose a final set of four. We can *not* say, however, that our final set of four is the best possible set from the original 26 features. As Cover and van Campenhout (9) have shown, only by exhaustively

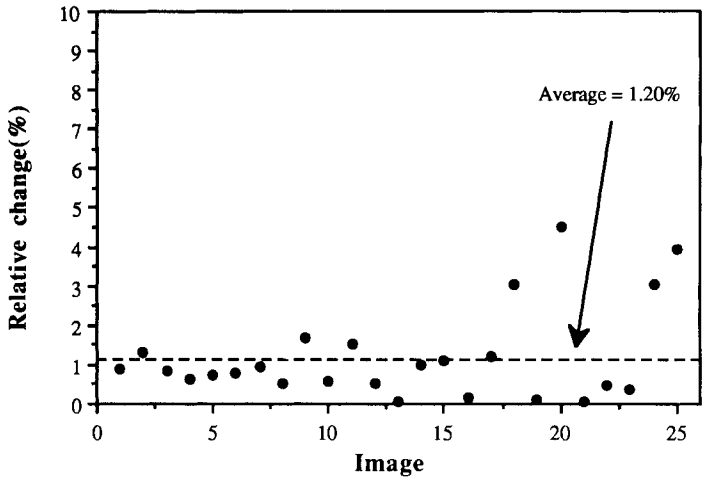


FIG. 7. Relative change per image when the same processing/analysis procedure is applied to 5 "scans" of a field-of-view. The percentage change in the linear combination of the four measured features is shown on the vertical axis. The average value of the relative change across all five sections (25 images) is 1.2%.

evaluating all $\binom{26}{4}$ = 14,950 possible combination of features can the best set of four be found. We preferred, however, not to surrender control of the choice of the four features to an exhaustive search procedure.

We feel that it is important to be able to check the features that have been selected at every step in the process for the presence of statistical outliers that might unduly influence an exhaustive procedure; MacSpin is ideally suited for this. As importantly, the features chosen must make sense in the context of tumor diagnosis. Clearly, a delicate balance must be struck between the acceptance of new, objective measures of cell and tumor pathology and the rejection of measures that, on the face of it, are irrelevant to the state of the tumor.

Reproducibility

Measurements can only be used to support a decision when they are reproducible. Reproducibility can be low because the measurements are disturbed by noise (random variations), because of instrument instabilities (deterministic variations), and because the phenomenon itself may exhibit natural heterogeneity (biological variations). The reproducibility of our measurements is examined in this context.

Random variations. Five images were acquired *at the same location* from each of five different sections of various grades for a total of 25 images. Reasons for differences between the five copies might be electronic noise, vibrations in the microscope table, fluctuation of the illumination source, and/or variations in the exposure time. All of these are stochastic (that is, random) in nature. The "standard procedure" (segmentation, binary processing, and measurement) was applied to these images and the linear combination $\mathbf{q}$ of the four features derived for each of the 25 images. For the five images taken from a given section we can define an average value of $\mathbf{q}$ as $\mathbf{m_q}$ (where the average is taken over the five acquired images) and a relative change $\epsilon_i(\%)$ as

$$\epsilon_i(\%) = \frac{|\mathbf{q_i} - \mathbf{m_q}|}{\mathbf{m_q}} \cdot 100\% \qquad (i = 1,2,3,4,5), \tag{3}$$

where $\mathbf{q_i}$ is the linear combination for the ith image of the same section at the same location. The results of this experiment are given in Figure 7. This figure demonstrates that the results of the overall measurement procedure are quite reproducible. The average variation was less than 2% for the linear combination of features used in this study.

Deterministic variations. The images used in this project have been acquired using a high-quality, low-noise camera (≥ 58 db.) and a relatively sophisticated microscope system. It requires accurate instrument handling and instrument set-up to acquire images that are exactly in focus. It is difficult to guarantee that the image seen through the user's eyepiece is exactly parfocal with the image projected (via the camera's eyepiece) onto the camera sensor. For this reason we investigated the behavior of the measurement procedure when a certain level of defocus is simulated.

Defocussing can be seen as a convolution of the image with a "pillbox." A uniform linear filter can be seen as an approximation to this convolution. By using several filter sizes (5 × 5, 7 × 7, 9 × 9, and 11 × 11), different levels of defocussing can be simulated. Applying a uniform filter of size 11 × 11 results in a very unsharp image; it is unrealistic to suppose that images would be worse than that.

Each of these filters was applied to a given image and the "standard procedure" then applied. The resulting value of $\mathbf{q}$ was derived for each image and com-

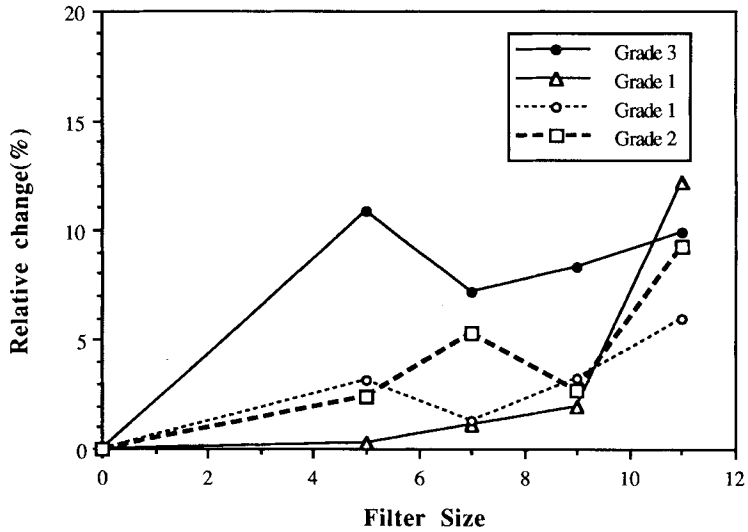


FIG. 8. Relative change per image when the same processing/analysis procedure is applied to images of tumors of different grades that have been artificially defocussed.

pared to the "true" value from the image without the simulated defocussing. Again a relative change $\epsilon(\%)$ was computed but this time using the single "in focus" image as $\mathbf{m}_q$. The results of this experiment are given in Figure 8.

The first conclusion that can be drawn is that measurements in defocussed images are possible. The consequences of the defocussing can also be analyzed. Differences between objects are smoothed away as a result of the simulated defocus. Because the linear combination contains three CVs, this effect tends to reduce the value $\mathbf{q}$ of the linear combination. After spatial smoothing, objects that are close together are frequently seen as one object by the segmentation procedure. This results in a double perimeter. Because the mean perimeter is one of the four features of the linear combination this effect will increase the value of the linear combination. The net result is the sum of the two effects.

Biological variations. Nine different images were collected from 19 tissue sections. The microscope fields-of-view were selected by an independent, *inexperienced* observer from the region that had been indicated by a pathologist as a characteristic for the tumor. The tumor grade was kept unknown from the observer during the selection of the microscopic field. The use of an inexperienced observer was intended to stimulate the choice of a random field-of-view from within a tumor section.

Each of the 19 sections was represented by the measurements in *three* images, compressed into *one* linear combination of four features as previously described. The nine images per section therefore resulted in *three* independent values of the linear combination per section. The variations in the values of the linear combination for the same section are considerable as can be seen in Figure 9.

It is easy to see in Figure 9 that the difference in the objective measure generated by the linear combination of features varies far more than the average of 2% that could be attributed to random sources of variation. Further, since we tried to work exclusively with well-focussed images, the variation can not be attributed to what we have termed deterministic variations.

CONCLUSIONS

Segmentation

The segmentation procedure that was developed yielded acceptable results. Because it is based on *local* contrast it can cope with a number of problems that are imposed by the structure of tissue sections. We did not require, nor was the segmentation procedure capable of, the successful segmentation of each and every cell nucleus within a tissue section. We accepted the basic idea that some touching and overlapping nuclei would never be separable. The question then becomes: Can an objective measure for tumor grading be possible when measurements are performed on segmented objects where the objects consist of one or more cell nuclei? This question has been addressed in the work of Bacus (2). He concluded that in the practice of pathology it is not the individual cell that "displays" the diagnostic information. Rather it is the cell in its proper context, surrounded by and touching other cells.

Evaluation of features

Size and shape. This study has shown that measurable size and shape differences for the cellular objects in tumors of different grades exist. Both the mean values and the Coefficients of Variation (CVs) offer dis-

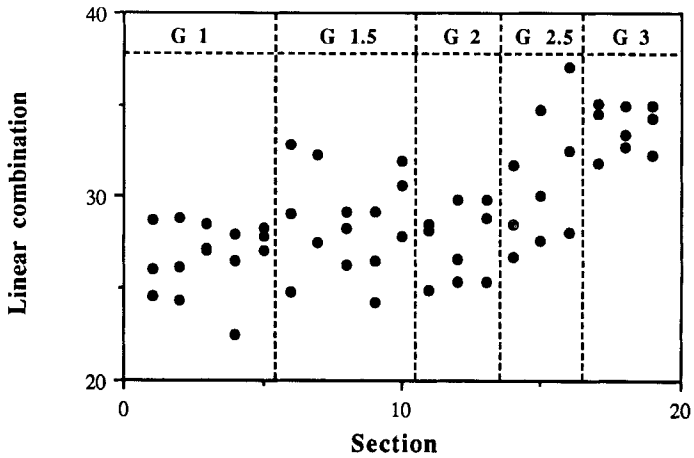


FIG. 9. Reproducibility of measurements taken from 19 different sections ranging from grade 1 to grade 3. The linear combination of the four measured features is shown on the vertical axis. Each symbol (•) represents the combined data from three images. For each of the 19 sections there are three (vertically arranged) symbols showing the spread of measurements within each section.

criminatory capability. The CV of a size or shape feature can be seen as a feature describing tissue architecture. According to these size and shape differences the classes grade 1 and grade 3 are clearly distinct; grade 2 tumors form a large intermediate zone between grade 1 and grade 3. The learning set is not large enough to ascertain exactly which set of size and shape features discriminates best between the three classes. The results indicate that this set most likely contains a feature describing *mean* size, a feature describing *variation* in size and a feature describing variation in *shape*.

DNA content. Because the segmentation and measurement procedure permitted multiple nuclei to be measured as one object, the optical density of the objects could not be used to estimate the distribution of the DNA in the cells.

Choice of stain. We chose the Feulgen procedure for this study in order to incorporate the possibility of measurements of stain content, that is, the stoichiometric link from optical density to stain content to DNA. For the reasons stated in the preceding paragraph, the DNA-based features proved ineffective. This leads us to the question/conclusion that the use of hematoxylin eosin stain might have been equally effective in this study.

Texture measures. With the instrumental set-up used in this research, the texture measures provided no discrimination between tumors of various grades. The use of too low a magnification and too low a sampling density might be a reason for this. Texture measures have been successful in cytologic studies where malignant cells must be distinguished from normal cells. Cells with a strange chromatin pattern are always suspicious. In tumor grading the appearance of the tissue *as a whole* is taken into account. Not *all* cells in a part of tissue section have a strange chromatin pattern. Therefore the discriminating power of the texture features may be low when the average over a large number of cells is taken.

Architecture measure. The architecture measure, based on nearest neighbor distances, tries to model the situation where, when tissue architecture is chaotic, some nuclei are close together and others are widely apart. The discriminating power of this feature was worse than many of the Acuity features discussed before. This may be explained from the fact that in many case two touching nuclei are treated as one object. As a result of this, variations in size and shape increase but variations in nearest neighbor distances can remain constant.

Reproducibility. It can be concluded that the reproducibility of the measurements performed by an inexperienced user is only moderate (see Fig. 8). The main cause of these large variations in the measurements is that the tumor tissue itself varies over the slide. Noise and instrument variations are not the cause of these large variations.

SUMMARY

As human beings we are inclined to divide a continuous space into a number of manageable, understandable categories. An automobile tachometer, capable of providing a smooth, quantitative measure of engine activity, is divided by the manufacturer into three regions: green, yellow, and red. This is presumably easier for the user to grasp.

The division of tumors into grades 1, 2, and 3 may be a similar example of this phenomenon; a categorization induced by "science" not nature. Our results show that there is a significant quantitative difference between grade "1" tumors and grade "3" tumors. The labels "1" and "3", however, should not be confused with actual numbers. While the labels may be ordered with respect

to the severity of the tumor stage, it is not true that "3" = 3·"1". What is much more likely is that the labels "1" and "3" are just two distinct points in a continuous process.

Our results have shown that the label "2" is not quantitatively distinct from "1" and "3." On an objective basis, however, we have shown that it lies somewhere between the other two stages. What is likely is that "2" is simply a reflection of our inclination to provide a third label between the two extremes, "1" and "3", when, in fact, there is only a continuous transition between them. This is supported by the fact that, as of this writing, we know of neither an immunological probe nor a cytogenetic probe that is capable of objectively distinguishing between the three classes.

The results, specifically those in Figure 6b, indicate that it is possible to assess tumor grade on a quantitative scale. The possible variations that diminish the reproducibility of these results, in particular that seen in Figure 9 and, to a lesser extent, in Figure 8, indicate that these measurements must not be seen as an automatic *substitute* for a pathologist but as a quantitative and semi-objective *support* for the decision-making process.

LITERATURE CITED

1. Abmayr W, Mannweiler E, Oesterle D, Deml E: Segmentation of scenes in tissue sections. Anal Quant Cytol Histol 9:190–196, 1987.
2. Bacus JV: Application of digital image processing techniques to cytology automation. Rush-Presbyterian-St. Luke's Medical Center, NCI Contract Report NO1-CB-74190, September 1982.
3. Blomjous CEM: Prognostic factors in urinary bladder carcinoma: A morphometric, DNA flow cytometric, immunohistochemical, and ultrastructural study. Ph.D. thesis, Faculty of Medicine, Free University of Amsterdam, 1988.
4. Blomjous CEM, Smeulders AWM, Baak JPA, Vos W, Van Galen CM, Meyer CJLM: A comparative study in morphometric grading of transitional cell carcinoma of the urinary bladder. Anal Quant Cytol Histol 11:426–432, 1989.
5. Chaudhuri BB, Rodenacker K, Burger G: Characterization and featuring of histological section images. Pattern Recognition Lett 7:245–252, 1988.
6. Chow CK, Kaneko T: Automatic boundary detection of the left ventricle from cineangiograms. Comp Biomed Res 5:388–410, 1972.
7. Clark PJ, Evans FC: Distance to nearest neighbor as a measure of spatial relationships in populations. Ecology 35:445–453, 1954.
8. Cooley WW, Lohnes PR: Multivariate Data Analysis. John Wiley and Sons, New York, 1971.
9. Cover TM, Van Campenhout JM: On the possible orderings in the measurement selection problem. IEEE Trans Syst Man Cyber SMC-7:657–661, 1977.
10. De Voogt HJ, Rathert P, Beyer-Boon ME: Urinary Cytology. Springer-Verlag, Berlin, 1977.
11. Fukunaga K: Introduction to Statistical Pattern Recognition. Academic Press, New York, 1972.
12. Gelsema ES: Ispahan User's Manual. Department of Medical Informatics, Erasmus University, Rotterdam, The Netherlands, 1988.
13. Giardina CR, Dougherty ER: Morphological Methods in Image and Signal Processing. Prentice-Hall, Englewood Cliffs, New Jersey, 1988.
14. Haralick RM, Shapiro L: Computer and Robot Vision. Addison-Wesley, Reading, Massachusetts, 1991.
15. Mostofi FK, Sobin LH, Torloni H: Histological typing of urinary bladder tumours. In: International Histological Classification of Tumors. World Health Organization, Geneva, 1973.
16. Piller H: Microscope Photometry. Springer-Verlag, Berlin, 1977.
17. Preston K: Interactive system for medical image processing. In: Real Time Medical Image Processing. Onoe M, Preston K, Rosenfeld A (eds). Plenum Press, New York, 1980, pp 193–206.
18. Remington RD, Schork MA: Statistics with Applications to the Biological and Health Sciences. Prentice-Hall, Englewood Cliffs, New Jersey, 1970.
19. Ridler TW, Calvard S: Picture thresholding using an iterative selection method. IEEE Trans Syst Man Cyber SMC-8:630–632, 1978.
20. Schmidt JD, Weinstein SH: Pitfalls in clinical staging of bladder tumors. 3:107, 1976.
21. Serra J: Image Analysis and Mathematical Morphology. Academic Press, London, 1982.
22. Sudar D: Quantitative characterization of tissue architecture. M.Sc. thesis, Faculty of Applied Physics, Delft University of Technology, 1988.
23. Tezuka F, Sato I, Higashiiwai H, Endo N, Ito K, Kasai M: Method for the quantitative evaluation of the distribution of nuclei in normal and malignant endometrial epithelia. Anal Quant Cytol Histol 12:237–241, 1990.
24. Van Vliet LJ, Young IT: ICAS: Scanner and analyzer for fluorescent microscopy. In: Abstracts of the XIV International Meeting of the Society for Analytical Cytology, Asheville, North Carolina, March 18–23, Cytometry Supplement 4:5, 1990.
25. Young IT: Sampling density and quantitative microscopy. Anal Quant Cytol Histol 10:269–275, 1988.
26. Young IT: Image fidelity: Characterizing the imaging transfer function. In: Fluorescence Microscopy of Living Cells in Culture: Quantitative Fluorescence Microscopy—Imaging and Spectroscopy, Taylor DL, Wang YL (eds). Academic Press, San Diego, 1989, pp 1–45.
27. Young IT, Roos R: Acuity: Image analysis for the personal computer. In: Pattern Recognition and Artificial Intelligence, Gelsema ES, Kanal LN (eds). Elsevier Science Publishers, Amsterdam, 1988, pp 5–16.
28. Young IT, Sudar D: Characterization of tissue architecture. In: Abstracts of the XIIth International Meeting of the Society for Analytical Cytology, Breckenridge, Colorado, Cytometry Supplement 2:571D, 1988.
29. Young IT, Verbeek PW, Mayall BH: Characterization of chromatin distribution in cell nuclei. Cytometry 7:467–474, 1986.