# Statistical Geometric Features — Extensions for Cytological Texture Analysis

Ross F. Walker Paul T. Jackway

Cooperative Research Centre for Sensor Signal and Information Processing
Department of Electrical and Computer Engineering, University of Queensland

E-mail: walker@elec.uq.edu.au http://www.cssip.elec.uq.edu.au/staff/walker.html

## Abstract

*Statistical Geometric Features (SGF) have recently been proposed for the classification of image textures. The SGF method is easily extended to use other geometric properties of connected regions. Following a brief review of the method, we propose such an extension to the set of SGF features for the purpose of classifying cervical cell textures. The resulting method proves to be as powerful as the Gray Level Co-occurrence Matrix (GLCM) method of texture analysis, when tested on a set of 117 cervical cell images. The ability to define features tailored to the geometric properties of the textures concerned makes this method a powerful analysis tool.*

## 1. Introduction

The extraction of descriptive features from image texture is an important task in image analysis and understanding, and an area of active research interest.

Recently, Chen, Nixon & Thomas have proposed a novel set of 16 features for texture classification called "statistical geometric features" (SGF)[1]. This work is of immediate interest since a thorough test on *all* 112 "Brodatz" textures has shown that the SGF method exhibits a "substantially higher" correct classification rate than three other current methods [6][10][14]. Further, the reduction in performance with an increasing number of textures is slower and the performance with additive noise is good [1].

The SGF approach is to decompose a gray-scale texture image into a stack of binary images via threshold decomposition. Certain geometric properties of the connected regions (foreground and background) in each binary image are then measured, and a number of statistical parameters based on these geometric properties are then computed. These parameters then become the extracted texture features for the purpose of texture classification.

The standard SGF method uses "number of connected regions" and "irregularity" as the only two regional geometric properties [1]. However, there are many other possible geometric properties of connected regions in binary images, therefore there is ample scope for the extension of the SGF method. Further, it should be possible to select the geometric properties to maximise classification performance in the particular texture problem at hand.

In this paper we explore such extensions to the SGF method for the discrimination of normal and abnormal cervical cell images by classifying their nuclear texture. We have previously shown that the Gray Level Co-occurrence Matrix (GLCM) method mentioned earlier performs quite well on the cervical cell texture problem [12, 13]. The GLCM method therefore becomes our benchmark to assess this new method.

We review the SGF method in the next section before introducing our proposed extensions in section 3. The new features as well as those of [1] are evaluated in a feature selection and cell classification methodology which we describe in section 4. The results are presented and discussed in section 5 and our conclusions appear in section 6.

## 2. Statistical Geometric Feature Algorithm

A discrete gray-scale image on a domain $\boldsymbol{D} \in \mathbb{Z}^2$ of $N_g$ gray levels is modelled as a 2D function $f : \boldsymbol{D} \rightarrow \boldsymbol{G}$, where the range $\boldsymbol{G} = \{0, 1, \ldots, N_g - 1\}$. The statistical geometric feature algorithm as given in [1] is:

**Step 1:** A stack of binary images $f_b(x, y; \tau)$ is produced from $f(x, y)$ by thresholding at each discrete intensity level $\tau \in \{1, 2, \ldots, N_g - 1\}$. For each binary image $f_b(x, y, \tau)$, a group of '1' valued pixels is defined as being a *4-connected region* if for all pixels in the group, each pixel has at least one 4-connected neighbour within the group. Groups of '0'-valued pixels are similarly defined.

**Step 2:** A geometric property is measured for each 4-connected region in each binary image. These measurements are then summed or averaged across all the '1'-valued regions and all the '0'-valued regions at each threshold to give a pair of geometric properties $g_1(\tau), g_0(\tau)$ as functions of threshold, $\tau$.

1015-4651/96 $5.00 © 1996 IEEE
*Proceedings of ICPR '96*

**Step 3:** Several statistics which characterise the distributions of $g(\tau)$ across $\tau$ are then computed. These statistics are then used as texture features for classification. ■

Two sets of geometric properties are used by [1]. The first is a simple count of the number of connected regions:

$$NC_k(\tau) = \text{number of 'k'-valued regions, } k = \{0, 1\}, \quad (1)$$

where $k = 0$ for '0'-valued regions and $k = 1$ for '1'-valued regions. The second, an average measure weighted by region size, of the irregularity or non-circularity of the regions, is defined as:

$$\overline{IRGL}_k(\tau) = \frac{\sum_{j=1}^{NC_k(\tau)} IRGL_j(\tau).NOP_j(\tau)}{\sum_{j=1}^{NC_k(\tau)} NOP_j(\tau)}; \quad (2)$$

where index $j$ is the $j$th 4-connected region, $NOP_j(\tau)$ is the number of pixels in the $j$th region at level $\tau$, and $IRGL_j(\tau)$ is the irregularity or non-circularity of each region, given by:

$$IRGL = \frac{1 + \sqrt{\pi}.\max_{i \in I} \sqrt{(x_i - \bar{x})^2 + (y_i - \bar{y})^2}}{\sqrt{|I|}} - 1, \quad (3)$$

where, $I$ is the set of all indices to pixels in the region, $|I|$ the number of indices (cardinality) of this set, and

$$\bar{x} = \sum_{i \in I} x_i/|I|, \quad \bar{y} = \sum_{i \in I} y_i/|I|. \quad (4)$$

The four feature functions of threshold level $\tau$ defined above ($NC_1(\tau)$, $NC_0(\tau)$, $\overline{IRGL}_1(\tau)$, $\overline{IRGL}_0(\tau)$) represent statistical distributions. [1] defines four statistics based on these feature functions, namely,

$$max\ value = \max_{\tau} g(\tau), \quad (5)$$

$$average\ value = \frac{1}{N_g - 1} \sum_{\tau} g(\tau), \quad (6)$$

$$sample\ mean\ \bar{\tau} = \frac{1}{\sum_{\tau} g(\tau)} \sum_{\tau} \tau.g(\tau), \quad (7)$$

$$sample\ S.D. = \sqrt{\frac{1}{\sum_{\tau} g(\tau)} \sum_{\tau} (\tau - \bar{\tau})^2.g(\tau)}, \quad (8)$$

where $g(\tau)$ is one of the four feature functions. This gives a total of 16 features based on the statistics of the geometric properties of the image.

## 3. Analysis of the Proposed Method

### 3.1. Cytological Representation of SGF Regions

Intensity images of cervical cell nuclei are a representation of chromatin *density* within the nuclei. Areas of condensed chromatin known as *heterochromatin* absorb larger quantities of stain than the more sparse *euchromatin*. Thus, low intensity areas of a nuclear image represent predominantly heterochromatin, while high intensity areas represent euchromatin. The use of threshold level $\tau$ effectively segments the nuclear image based on chromatin density. Figure 1 details a series of thresholded images of a single nucleus. Features based on '1'-valued pixels are measures of nuclear regions containing predominantly euchromatin. Features based on '0'-valued pixels measure characteristics of nuclear regions containing predominantly heterochromatin. For example, in cytological terms, $NC_1$ represents the number of euchromatin clumps, while $NC_0$ represents the number of heterochromatin (condensed chromatin) clumps. This representation is somewhat weaker at the extreme $\tau$ levels. For example, at threshold $\tau = 1$, '1'-valued clumps represent areas containing not only euchromatin, but also low density heterochromatin. At $\tau = n_g - 1$, '0'-valued regions represent areas containing all heterochromatin plus an amount of the higher density euchromatin. This presents a problem when analysing the results of subsequent feature classification in terms of heterochromatin or euchromatin properties. To overcome this problem, we calculate '1'-valued features at threshold levels $\tau = 4, \ldots, n_g - 1$ and '0'-valued features at threshold levels $\tau = 1, \ldots, n_g - 4$. This removes the more "contaminated" clumps from the analysis, allowing stronger conclusions to be drawn.

### 3.2. Refinements to Features

Ideally, texture measures should be independent of the amount or area of texture analysed, i.e. of window size. This is particularly important for applications where it is impractical or impossible to select a fixed window size (i.e. cell texture analysis). The feature $NC$ is unfortunately linearly dependent on image size. That is, doubling the image area with the same texture will double the number of connected regions in the image. Thus, the feature as proposed by Chen *et-al* when used for cell nuclei is not specifically a texture measure, but a measure of both texture *and* nuclear size.

We propose to re-express this feature in a form that is independent of image size, by normalizing the measure based on image area,

$$NCA(\tau) = NC(\tau)/|I_{f(x,y)}| \quad (9)$$

where $NCA$ is the number of connected regions normalized by the the image area, $I_{f(x,y)}$ is the set of pixel indices in the image $f(x, y)$, and $|I|$ is the cardinality of $I$.

### 3.3. New Features

The ability to define new features specific to the problem at hand represents a significant advantage of the SGF method. The use of such 'tailored' features not only allows better 'targeting' of possible discriminatory properties

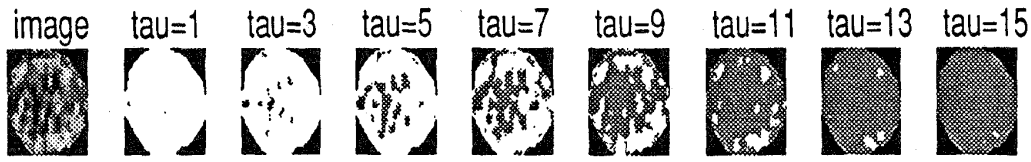


**Figure 1. A series of thresholded images of a single nucleus**

within a texture, but also allows much stronger conclusions to be drawn from subsequent classification results. Prior to feature definition, an analysis of texture properties and an understanding of the process which generated the texture is warranted.

The following features are an attempt to measure specific cytological properties of the heterochromatin and euchromatin clumps in cell nuclei (to be referred to subsequently as 'dark' and 'light' clumps, respectively). These measures are based on properties of the cell such as the centre of gravity of the nucleus, clump size, and clump position within the nucleus (contextual information). Such features have been reported to occur with dysplasia in cells [2, 8].

Define the centre of gravity of the $j$th clump at some threshold $\tau$ as $(\bar{x}_j, \bar{y}_j)$. Define the centre of gravity of the entire nucleus as

$$x_{COG} = \frac{\sum_{i \in I_N} x_i}{|I_N|}, \quad y_{COG} = \frac{\sum_{i \in I_N} y_i}{|I_N|} \tag{10}$$

where $I_N$ is the set of all indices to pixels in the entire nucleus, and $|I_N|$ is the cardinality of $I_N$.

Define the normalized clump displacement of the $j$th clump from the centre of gravity of the nucleus as

$$D_j = \sqrt{\pi}\frac{\sqrt{(\bar{x}_j - x_{COG})^2 + (\bar{y}_j - y_{COG})^2}}{\sqrt{|I_N|}}. \tag{11}$$

For circular regions, $D_j$ is expressed as a proportion of the radius of the region.

**Average Clump Displacement**

$$\overline{DISP}(\tau) = \sum_j D_j / NC(\tau) \tag{12}$$

Measures the average displacement of regions from the centre of gravity of the nucleus (normalized for nuclear area). This feature attempts to measure whether cell abnormality results in clumps whose displacements from the centre of gravity of the nucleus are, on average, greater or less than that of normal cells.

**Average Clump Inertia**

$$\overline{INERTIA}(\tau) = \sum_j D_j . NOP_j(\tau)/NC(\tau) \tag{13}$$

Measures the average inertia of regions, where *inertia* is defined as the product of clump area times clump displacement from center of gravity. This feature attempts to determine whether cell abnormality results in contextual changes in chromatin clump distribution. That is, whether larger chromatin clumps are displaced further from or closer to the nucleus centre of gravity.

**Total Clump Area**

$$TAREA(\tau) = \sum_j NOP_j(\tau)/|I_N| \tag{14}$$

Measures the total area of regions relative to the area of the nucleus. This feature will determine whether cell abnormality results in more/less chromatin as a proportion of cell area.

**Average Clump Area**

$$\overline{CAREA}(\tau) = \sum_j NOP_j(\tau)/NC(\tau) \tag{15}$$

Measures the mean area of clumps. Any correlation between cell abnormality and increased/decreased chromatin clump size will be detected by this feature.

## 4. Feature Evaluation

We now apply the 4 statistical measures (5)-(8) to the above 6 feature functions, giving a total of 24 feature measures for each of the two clump types.

### 4.1. Cell Database

The data consists of a set of 117 cells captured from 12 cervical slides processed using ThinPrep® slide preparation[1] and regular Papanicolaou staining. A total of 57 cells with abnormalities ranging from mild dysplasia (CIN1) to Carcinoma-in-situ (CIS) were captured from 11 abnormal slides, while 60 normal cells were captured from both the 11 abnormal and 1 normal slides. The imaging system was photometrically recalibrated between each capture session, and the imaging of normal and abnormal cells were randomly interspersed. All cells were classified through the

[1] Cytyc Corporation, Massachusetts, U.S.A.

microscope before imaging, by a Cytologist. Examples of typical normal and abnormal cell nuclei are shown in Figure 2. It can be seen from these images that it is quite difficult for the untrained observer to distinguish visual differences between normal and abnormal cell nuclei in isolation.

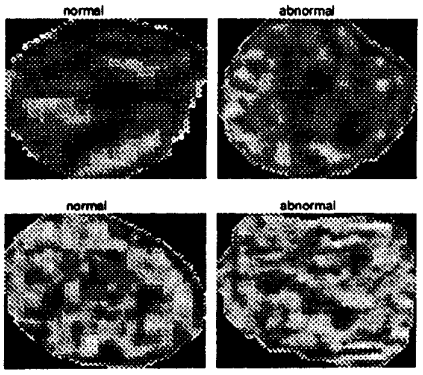


**Figure 2. Typical examples of both normal and abnormal cervical cell nuclei, segmented using the technique described in subsection 4.2**

### 4.2. Nuclear Segmentation

Following cell imaging and capture, each nucleus was segmented from its surrounding cytoplasm using a series of automated fast morphological transforms with octagonal structuring elements[9]. Full details can be found in [13]. The resulting images were requantized to 16 gray levels to reduce the computational expense.

### 4.3. Feature Preprocessing

The class separability measure and classifier used in the following work are based on parametric methods, where Gaussian distributed data is assumed. Because the statistical distribution of the defined SGF features is unknown, we pre-process all features before feature selection and classification. We re-express all features using the *Ladder of Powers* technique[11], further details can be found in [13].

### 4.4. Feature Selection and Classification

Discriminant Analysis is used to reduce the high dimensionality of the feature space to a lower dimension to better characterize the class distributions within this new space. This is achieved by removing redundant features which results in a remaining feature set with maintained discriminatory power. Based on the total of 117 cell patterns, and the rule-of-thumb of 3 patterns/feature/class[3], we reduce the 48 dimensional feature space to 20 dimensions via discriminant analysis, using the Bhattacharyya discrimination measure $J_B$[5] and Kittler's *Plus 2-Take Away 1* feature-set search algorithm[5, 7]. As a comparison of the SGF technique to that of the GLCM method, we also present the results of classification of cell texture for optimised SGF and GLCM feature sets from 1 to 20 dimensions. A total of 40 GLCM features were trialed, derived from 8 standard feature functions (*Energy, Entropy, IDM, Correlation, Inertia, Variance, Shade, Promenance*) at displacements of 1,2,4,8,and 16 pixels. Further details can be found in [12, 13].

Classification was performed by a general Bayes decision function for Gaussian feature distributions with unequal variance-covariance matrices[4]. The resulting decision boundary is of hyperquadric form. Leave-one-out classification was implemented to provide accurate estimation of the real classification error, based on the small data set of 117 cells. For each trial, the quadratic classifier was trained on all but one sample, and the performance of the resulting classifier evaluated on this sample. This process was repeated until all samples had been classified once. The sum of the misclassifications represents the real (as opposed to the apparent) misclassification rate, and is an unbiased estimate.

## 5. Results

### 5.1. Discriminant Analysis Results

Figure 3 details the real misclassification rates produced by leave-one-out classification. As a comparison, the results for Gray Level Co-occurrence Matrix features from previous work [13] are also shown.

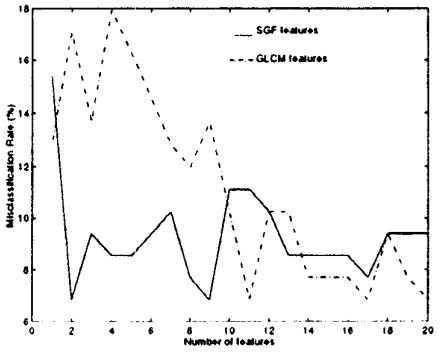


**Figure 3. Comparison of SGF and GLCM misclassification rates for optimal feature sets.**

It can be seen that SGF features provide better classification performance than GLCM features at lower dimensions. This may be due to the tailoring of SGF features to measure

specific texture properties, as opposed to the more *ad-hoc* application of pre-defined GLCM features. It can also be seen that most of the discriminatory power exhibited by the SGF features appears to be contained in two features.

### 5.2. Feature Distribution Analysis

Figure 4 details a number of features which appear to capture some chromatin changes during cell dysplasia.

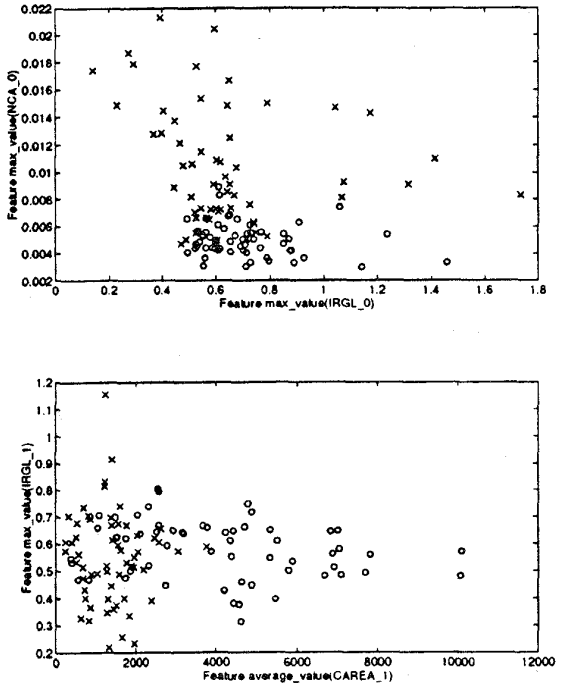


**Figure 4. Features which exhibit distribution changes during cell dysplasia.**

From these two plots it can be inferred that:

Normal cells have more heterochromatin clumps per unit area then abnormal cells (feature *max_value(NCA_0)*).

The average size of heterochromatin clumps in abnormal cells is greater than that of normal cells (feature *average_value(CAREA_1)*).

The irregularity of abnormal cell chromatin exhibits far less variability than that of normal cells (feature *max_value(IRGL_0)*).

## 6. Conclusions

The method of SGF texture analysis using the features defined in this work, provides good discriminatory power when detecting textual changes in high-resolution cervical cell images. Preliminary results indicate that the method may be more powerful than the Gray Level Co-Occurrence Method [6]. Moreover, the use of feature functions derived specifically for the purpose of cell analysis allow quantitative as well as qualitative descriptions of chromatin texture changes in abnormal cell nuclei. This ability to define SGF features tailored to the geometric properties of the textures allows far stronger conclusions to be drawn from the feature distributions and classification results than is possible with many other texture methods, thus making this method a powerful analysis tool.

## References


[1] Y. Q. Chen, M. S. Nixon, and D. W. Thomas. Statistical geometric features for texture classification. *Pattern Recognition*, 28(4):537–552, 1995.

[2] H. E. Danielsen, G. Farrants, and A. Ruth. Characterization of chromatin structure by image analysis - a method for the assessment of changes in chromatin organization. *Scanning Microscopy Supplement*, 3:297–302, 1989.

[3] D. H. Foley. Considerations of sample and feature size. *IEEE Transactions on Information Theory*, 5:618–626, 1972.

[4] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Addison-Wesley Publishing, U.S.A., 1992.

[5] D. J. Hand. *Discrimination and Classification*. John Wiley and Sons, USA, 1981.

[6] R. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3:610–621, 1973.

[7] J. Kittler. Feature selection and extraction. In T. Y. Young and K.-S. Fu, editors, *Handbook of Pattern Recognition and Image Processing*, pages 60–83. Acedamic Press, San Diego, 1986.

[8] D. Komitowski and C. Janson. Quantitative features of chromatin structure in the prognosis of breast cancer. *Cancer*, 65(12):2725–2730, 1990.

[9] Y. H. Lee. Algorithms for mathematical morphological operations with flat top structuring elements. *SPIE Applications of Digital Image Processing*, 8:33–45, 1982.

[10] S. Liu and M. Jernigan. Texture analysis and discrimination in additive noise. *Computer Vision, Graphics, and Image Processing*, 49:52–67, 1990.

[11] P. F. Velleman and D. C. Hoaglin. *Applications, Basics, and Computing of Exploratory Data Analysis*. Duxbury Press, Boston, Mass., 1981.

[12] R. F. Walker, P. Jackway, and B. Lovell. Cervical cell classification via co-occurrence and markov random field features. In *Proceedings of DICTA-95, Digital Image Computing: Techniques and Applications*, pages 294–299, Brisbane, Australia, 30 Nov—2Dec 1995. IEEE.

[13] R. F. Walker, P. Jackway, B. Lovell, and I. Longstaff. Classification of cervical cell nuclei using morphological segmentation and textural feature extraction. In *Proceedings Second Australian and New Zealand Conference on Intelligent Information Systems*, pages 297–301, Brisbane, Australia, 30 Nov — 2 Dec 1994. IEEE.

[14] C. Wu and Y. Chen. Statistical feature matrix for texture analysis. *Computer Vision, Graphics, and Image Processing*, 54:407–419, 1992.