Characterization of Chromatin Texture by Contour Complexity for Cancer Cell Classification

Tomoharu Kiyuna, Akira Saito, Elizabeth Kerr, and Wendy Bickmore

***Abstract*— The purpose of this study is to investigate a new technique for image-based cancer cell classification and provide a more quantitative and objective characterization method for a diagnosis, which currently relies on qualitative and empirical judgment of pathologists. For this, a new method for chromatin texture characterization employing a new feature, *contour complexity*, is proposed and evaluated using nuclear images obtained from paraffin-wax embedded sections of human breast cancer on slides. The proposed feature is calculated on the basis of a contour length of nucleus obtained by setting different threshold values of intensity for a grayscale image, and it is a quantitative measure of chromatin texture. An expectation-maximization (EM) algorithm-based segmentation and an effective initial parameter search method for EM are used for the automatic calculation of the feature. The results for breast cancer cell detection showed that the average contour complexity value for malignant cells (19.6±4.1) is found to be significantly greater ($p < 10^{-6}$, Kolmogorov-Smirnov test) than that of benign cells (0.35±0.17). By the comparison with the conventional fractal dimension approach, it is shown that the proposed feature is much more sensitive feature than the fractal dimension for the individual cancer cell detection.**

I. INTRODUCTION

In spite of the recent progress in understanding the properties of cancer cells from a molecular biological point of view [1], the final diagnosis of cancer still relies on pathologists' judgment, which is less objective and difficult to quantify. Therefore, it has been expected that quantitative and objective characterization of cancer cells will promote more accurate evidence-based diagnosis.

A number of methods have been proposed to characterize cancer cells on the basis of digital images. The most obvious and important feature of a cancer cell is the nuclear size. However, some types of cancer, such as invasive lobular carcinoma and lobular carcinoma in situ, have relatively small nuclear sizes, and thus it is difficult to diagnose such types of cancers on the basis of nuclear size only.

Another important feature is chromatin texture. According to the degree of DNA condensation, chromatin can be categorized into two main classes, euchromatin and heterochromatin, and coarse and asymmetric aggregates of heterochromatin are observed throughout a wide range of cancers [2]. It is therefore important to characterize the chromatin texture to achieve a quantitative and objective cancer diagnosis.

Since there are many ways to characterize the texture in cancer diagnosis, a variety of methods for achieving this have been proposed, such as the gray level co-occurrence matrices [3], chromatin distributions [4], entropy [5], and fractal dimension [6]–[8].

Among these methods, the fractal dimension is particularly widely used in the field of pathology (see [9] for comprehensive reviews). Fractal dimension can be used to describe the self-similarity, the complexity, or the irregularity of objects. Although fractal dimension is widely used for characterizing cancer cells, it is not sensitive enough to detect individual cancer cells, and predict the histopathological diagnosis with sufficient accuracy [10].

The purpose of this study is to investigate a new technique for image-based cancer cell classification and provide a more quantitative and objective characterization method for the current pathology which relies on the qualitative and empirical judgment of pathologists. We propose a new method for chromatin texture characterization employing a new feature, *contour complexity*. The expectation-maximization (EM) based segmentation method and an effective parameter initialization method for EM are employed to calculate this feature automatically. The effectiveness of the proposed method is evaluated and compared with a fractal dimension based approach.

II. MATERIALS AND METHODS

A. Fluorescence in situ Hybridization (FISH)

Paraffin-wax embedded sections of human breast cancer on slides were heated to 60°C for 30 mins and washed 4 times in xylene for 10 mins each before rehydration through an ethanol series. They were then microwaved for 50 mins in 0.1M citrate pH6 buffer, washed and stored in water. The slides were washed for 40-50 mins in a pepsin/HCl solution at 37°C, rinsed in PBS then 2xSSC and dehydrated through an ethanol series and air dried.

The slides were incubated in 2xSSC for 5 mins at 75°C, denatured for 3 mins at 75°C in 70% formamide-2xSSC, plunged into ice-cold 70% ethanol for 3 mins, dehydrated through an ethanol series and air dried. Whole chromosome FISH was carried out using ready-made probes from Cambio (Cambridge, UK) according to the manufacturer's instructions. Briefly, the probe is hybridized to section overnight at 37°C, then washed 4 times for 3 mins in 2xSSC at 45°C, 4 times for 3 mins in 0.1xSSC at 60°C and stored

Manuscript received June 13, 2008; revised August 15, 2008. T. Kiyuna and A. Saito are with Biomedical Imaging and Informatics Group, Innovative Solutions Promotion Division, NEC Corporation, Shiba-koen First Building 15F, 3-8-2, Shiba, Minato-ku, Tokyo 105-0014, Japan (phone:+81-3-5441-6085; fax: +81-3-5441-6092; e-mail: {t-kiyuna@bl, a-saito@ak}.jp.nec.com).

E. Kerr and W. Bickmore are with the MRC Human Genetics Unit, Chromosomes and Gene Expression Section, Western General Hospital, Edinburgh, Crewe Road, Edinburgh, EH4 2XU, UK (e-mail:{elkerr, W.Bickmore}@hgu.mrc.ac.uk).

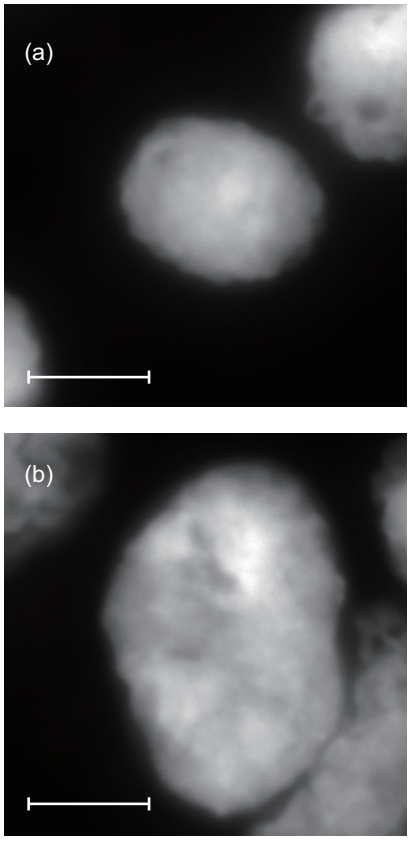

Fig. 1. Examples of input image. (a) benign cell; (b) malignant cell. Scale bars, 5μm.

in 4xSSC/0.01%Tween. Biotin labeled paints were detected using fluorochrome-conjugated avidin (FITC or Texas Red) (Vector Laboratories) followed by biotinylated anti-avidin (Vector Labs) and a final layer of fluorochrome-conjugated avidin. Digoxygenin labeled probes were detected using rhodamine anti-dig (Vector Labs) followed by a layer of Texas Red anti-sheep (Vector Labs). Slides were counterstained with 0.5μg/ml DAPI/vectashield.

B. Image Acquisition

FISH images were captured (256x256x24bit) at a magnification of 100X, using a Coolsnap HQ CCD camera (Photometrics Ltd, Tucson, AZ) attached to a Zeiss Axioplan II fluorescence microscope with 100x plan-neofluar oil objective, a 100W Hg source (Carl Zeiss, Welwyn Garden City, UK) and a Chroma #83000 triple band pass filter set (Chroma Technology Corp., Rockingham, VT) with the excitation filters installed in a motorised filter wheel (Prior Scientific Instruments, Cambridge, UK). Image capture and analysis were performed using in-house scripts written for IPLab Spectrum (Scanalytics Corp, Fairfax, VA). DAPI-stained nuclear images were obtained using blue channel of the original RGB images and saved in TIFF format (256x256x8bit). Fig. 1 shows examples of images used for analysis.

Algorithm 1 Initial Parameters for EM

Inputs k_0: `# of classes`, r_0: `initial cluster resolution`, $\{x_i, p_i\}_{i=1}^m$: `empirical density`, ε: `shrink rate`
$r \leftarrow r_0$ // `initialize cluster resolution`
while $k \leq k_0$ **do**
  $k \leftarrow 0$
  $I \leftarrow \{1, 2, \ldots, m\}$ // `indices of non-empty bins`
  repeat
    $j \leftarrow \operatorname{argmax}_{i \in I} p_i$
    $\mathcal{N}_j \leftarrow \{i \in I : \|x_j - x_i\| \leq r\}$ // r-`neighbor of` x_j
    if $\min_{i<k} d(x_j, C_i) < r$ **then**
      $C_i \leftarrow C_i \cup \mathcal{N}_j$ // `merge into overlapped cluster`
    else
      $k \leftarrow k + 1$
      $C_k \leftarrow \mathcal{N}_j$ // `accept as a new cluster`
    end if
    $I \leftarrow I \backslash \mathcal{N}_j$ // `remove processed indices`
  until $I = \emptyset$
  $r \leftarrow (1 - \varepsilon) r$ // `change cluster resolution`
end while
calculate $w_k, \mu_k,$ and σ_k using C_k $(k = 1, 2, \ldots, k_0)$

III. Statistical Image Analysis

A. Segmentation

Since our quantification method relies on the contour length, it is crucial to extract the nuclear area accurately and stably. For this purpose, we used EM-based image segmentation and parameter initialization for EM.

As a probabilistic model for a grayscale image, we used a finite normal mixture with k_0 components ($k_0 = 3$):

$$p(x_j) = \sum_{i=1}^{k_0} w_i f(x_j|\mu_i, \sigma_i) \quad (1)$$

$$= \sum_{i=1}^{k_0} w_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(x_j - \mu_i)^2}{2\sigma_i^2}\right], \quad (2)$$

where x_j is an intensity value at the j-th pixel, $f(x_j|\mu_i, \sigma_i)$ is the normal density of the i-th component with mean μ_i and standard deviation σ_i, and w_i is a mixture ratio. The parameters (i.e. mixture ratio, intensity mean and standard deviation) were estimated so as to maximize the likelihood using the EM algorithm [11]–[13]. Each pixel in the grayscale image was classified based on a posterior probability π_{ij} calculated as:

$$\pi_{ij} = \frac{w_i f(x_j|\mu_i, \sigma_i)}{\sum_{\ell=1}^{k_0} w_\ell f(x_j|\mu_\ell, \sigma_\ell)}. \quad (3)$$

That is, if $\pi_{\ell j}$ was the largest posterior probability out of π_{ij} $(i = 1, 2, \ldots, k_0)$ then the j-th pixel was classified as a member of the ℓ-th component. Pixels belonging to the component with the smallest mean intensity were segmented

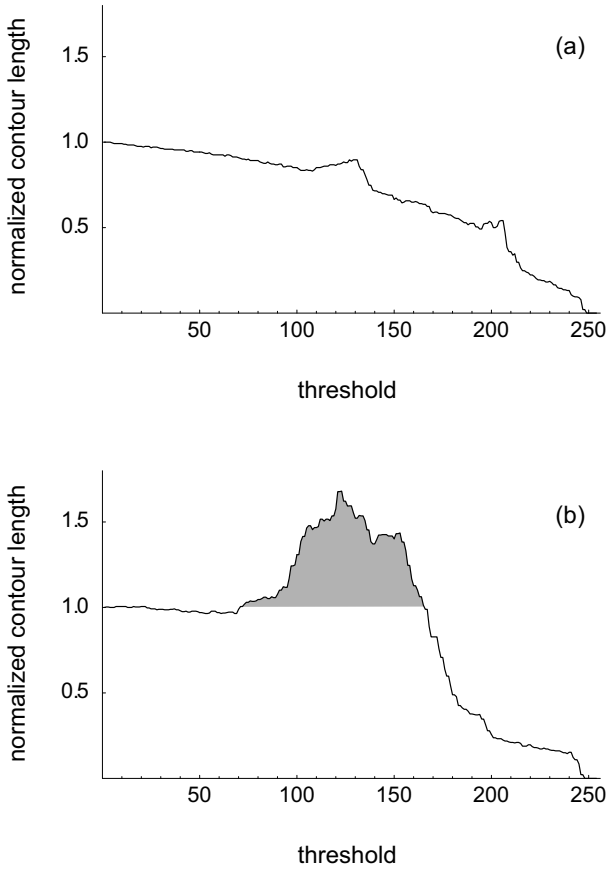

Fig. 2. Examples of contour length sequence obtained from (a) benign cell shown in Fig. 1a and (b) malignant cell shown in Fig. 1b. The gray area in (b) corresponds to contour complexity value.

as a background area, and the remaining pixels were regarded as a nuclear area.

To obtain a smooth boundary, a Gaussian filter was applied to the boundary pixels and their neighboring pixels. The boundary was then re-segmented using the threshold determined by the procedure explained above. Gaps or holes are filled when they exist in the nuclear area.

We refer to the contour length of an extracted nuclear area as "standard contour length", L^*. Smoothing and gap filling were processed only for the calculation of the standard contour length L^*.

B. Initial Parameters for EM Algorithm

Although the convergence of the EM algorithm is theoretically guaranteed, the parameters estimated with it are not necessarily globally optimal. When the initial parameters are not appropriate, the parameter values can become mired in a local maximum during the likelihood maximization.

Algorithm 1 shows a parameter initialization algorithm based on empirical density (i.e, normalized histogram).

This algorithm generates a sub cluster by merging bins in the r-neighbor of the peak density value, and the next sub-cluster is either merged into it or accepted as the second cluster depending on the overlap between sub clusters. With this algorithm it is possible to determine the initial parameters for EM automatically.

Algorithm 2 Contour Complexity

1: extract nuclear area
2: calculate standard contour length L^*
3: normalize maximum intensity in the nuclear area to 255
4: set threshold $i \leftarrow 0$
5: **while** $i \leq 255$ **do**
6: extract areas with intensity $\geq i$
7: calculate total contour length L_i of the extracted areas
8: $i \leftarrow i + 1$
9: **end while**
10: Calculate contour complexity (8) based on obtained contour length sequence $\{L_0, L_1, \ldots, L_{255}\}$

In the Algorithm 1, m is the number of non-empty bins in the empirical density, x_i and p_i are center value and probability mass of the i-th bin, respectively, and $\mathcal{N}_j$ is a set of bin indices in the r-neighbor of the j-th bin. Overlap $d(x, A)$ is defined as

$$d(x, A) = \min\{|x - x_i|, i \in A\}, \tag{4}$$

where A is a set of bin indices. When cluster resolution r is too large, all sub-clusters may be merged into a single cluster or result in a number of clusters less than k_0 . In both cases, the cluster resolution r is shrunk until the specified number of clusters, i.e., k_0, is obtained.

When k_0 clusters are successfully obtained, the initial parameters are calculated as

$$w_k = \sum_{i \in C_k} p_i, \tag{5}$$

$$\mu_k = \frac{1}{w_k} \sum_{i \in C_k} p_i x_i, \tag{6}$$

$$\sigma_k^2 = \frac{1}{w_k} \sum_{i \in C_k} p_i (x_i - \mu_k)^2. \tag{7}$$

Where C_k $(k = 1, 2, \ldots, k_0)$ is a set of bin indices in the k-th cluster.

C. Contour Complexity

Algorithm 2 shows how to calculate the contour complexity based on a grayscale image. The standard contour length L^* is defined as the contour length of the nuclear area extracted by the segmentation method described in Section III-A.

Using contour sequence $\{L_0, L_1, \ldots, L_{255}\}$, the contour complexity is defined as

$$F_{cc} = \sum_i \left[\left(\frac{L_i}{L^*} \right) - 1 \right] \mathbb{I}_{\{L_i > L^*\}}, \tag{8}$$

where L_i is the total contour length when the threshold is i, and L^* is standard contour length. The symbol $\mathbb{I}_{\mathcal{A}}$ takes a value of 1 when condition $\mathcal{A}$ holds, and 0 otherwise.

We call this feature, which is calculated from the sequence of contour lengths, *contour complexity*. The contour complexity is the quantization of chromatin texture unevenness.

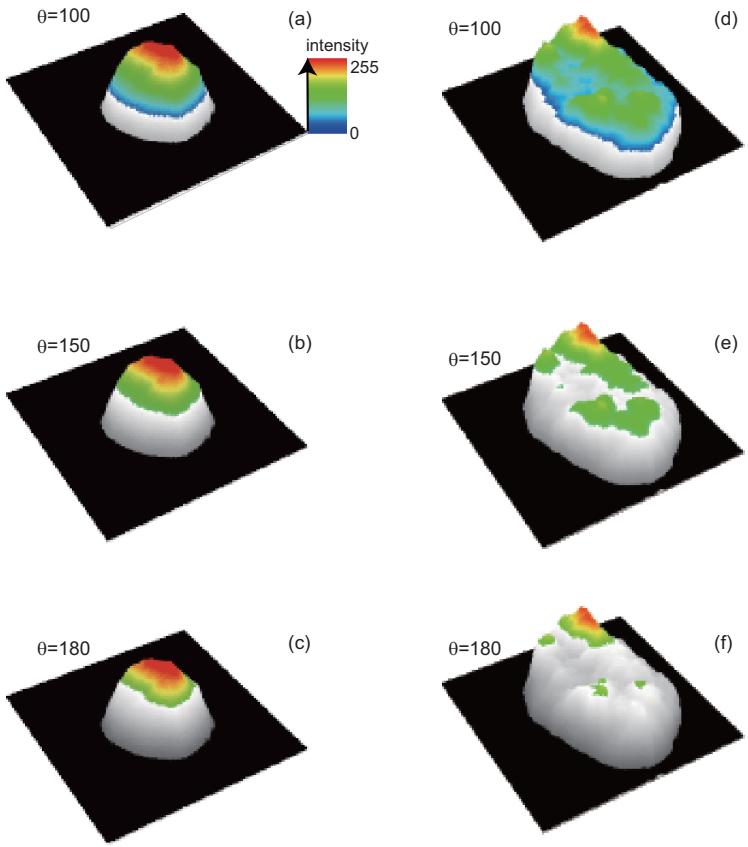

Fig. 3. Three-dimensional visualization of contour sequence. (a) to (c) benign cell; (d) to (f) malignant cell. The edge of area painted in color (blue to red) indicates the contour corresponding to the different thresholds. Thresholds (θ) are indicated in the figure.

This feature is based on the observation that when the degree of chromatin unevenness is high, many small clusters tend to be obtained in the nuclear area by thresholding, and the total contour length of these small clusters becomes longer.

Since contour length also becomes longer as the nuclear size becomes large, we normalized each L_i by dividing it by L^* so that the contour complexity is independent of nuclear size. In this way, we can evaluate the effectiveness of the contour complexity alone without the influence of nuclear size, which is also an important feature of cancer cells.

Fig. 2 shows examples of contour length sequence. The contour complexity value corresponds to the gray area in Fig. 2b. A three-dimensional visualization of the contour sequence is shown in Fig. 3. It is clear from this figure that when the degree of chromatin texture unevenness is high, the total contour length tends to become longer (e.g., compare Fig. 3b and Fig. 3e).

D. Fractal Dimension

To compare the validity of our method with that of previously employed methods, we conducted fractal analysis for the same samples. As a fractal feature, we used the maximum value of the fractal dimensions

$$D = \max\{D_i \mid i = 0, 1, \ldots, 255\}, \tag{9}$$

where D_i is the fractal dimensions obtained at the threshold i. The box-counting method is used to calculate the fractal dimension at each threshold.

IV. RESULTS AND DISCUSSION

To evaluate the effectiveness of the proposed feature for cancer cell classification, we compared the values of contour complexity between benign cells ($n = 30$) and malignant cells ($n = 37$). These cells were collected from paraffin-wax embedded sections of human breast tissue (two malignant tissues and two benign tissues).

Table I shows the statistics and Fig. 4 shows the distribution of the contour complexity for benign and malignant cells, respectively. It is clear from the table that the contour complexity of the malignant cells is significantly greater than that of benign cells ($p < 10^{-6}$, Kolmogorov-Smirnov test)[1]. No benign cells had a contour complexity greater than 5.0 and there were 10 cells with zero contour complexity. On the other hand, a majority of the malignant cells (21 out of 37) had a contour complexity greater than 5.0.

For the comparison, we applied the fractal dimension analysis [6] to the same samples. Table II and Fig. 5 show

[1]The Kolmogorov-Smirnov test is a non-parametric test to determine if two sample distributions are identical (can be used when the distributions are unknown). The very low score of 10^{-6} confirms that these two distributions are not identical.

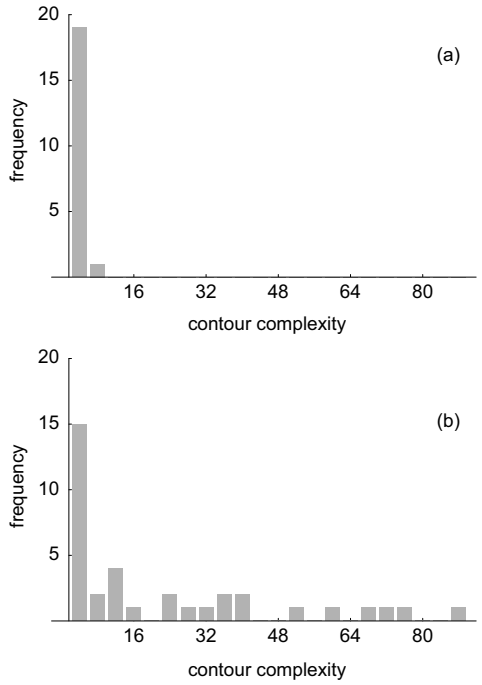

Fig. 4. Distribution of contour complexity. (a) benign cells; (b) malignant cells.

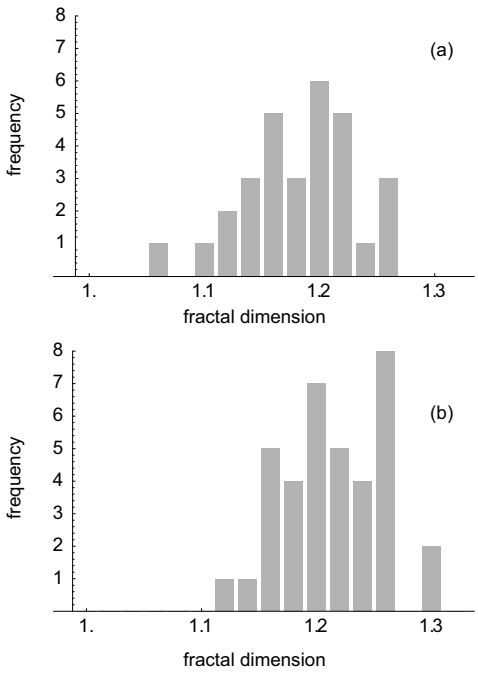

Fig. 5. Distribution of fractal dimensions. (a) benign cells; (b) malignant cells.

TABLE I

STATISTICS OF CONTOUR COMPLEXITY

	benign ($n = 30$)	**malignant** ($n = 37$)
mean	0.35±0.17	19.6±4.1
maximum	4.6	84.9
minimum	0	0

TABLE II

STATISTICS OF FRACTAL DIMENSION

	benign ($n = 30$)	**malignant** ($n = 37$)
mean	1.19±0.009	1.225±0.0072
maximum	1.26	1.30
minimum	1.07	1.13

the statistics and distributions of fractal dimensions. The fractal dimension of malignant cells was significantly larger than that of benign cells ($p < 0.01$, Welch's t-test)[2], which is consistent with the results obtained with our method. However, as shown in Fig. 5, it is difficult to distinguish malignant cells from benign cells from the fractal dimension only. This comparison showed that contour complexity is a more sensitive feature for chromatin-texture-based cancer cell classification than the fractal feature.

Fig. 6 shows examples of cells with low and high contour complexity values. These cells were collected from the malignant tissues. Although cells with the contour complexity less than 5.0 (Fig. 6a, Fig. 6b, and Fig. 6c) were collected from the malignant tissue, they do not necessarily appear to be malignant from the point of view of chromatin texture. In that sense, the values obtained with our proposed method are consistent with our intuitive judgment. The contour complexity value becomes especially greater when nucleoli appear in the nuclear area (Fig. 6d). The existence of nucleoli is also one of the important features of malignant cells. Here, too, we can see that the results obtained with our method are consistent with what our intuitive judgment tells us.

Although both contour complexity and fractal dimension features have been found to be effective means of distinguishing malignant and benign cells, the correlation coefficient (ρ) between two features is not so high (ρ=0.29 and 0.32 for malignant and benign cells, respectively). It is therefore considered that both features capture different aspects of chromatin texture.

With the EM-based segmentation method and the parameter initialization method described in the Sections III-A and III-B, the appropriate thresholds for the nuclear area extraction, which ranged from 30 to 98 in grayscale value, are adaptively determined and the contour complexity for 67 images are calculated in a fully automatic way without human intervention. The parameter initialization algorithm takes no longer than 1.0 second[3] in all images which is much faster than existing optimization methods such as a genetic algorithm and simulated annealing.

The results suggest that contour complexity is much more sensitive than the fractal dimension feature for the detection of individual cancer cells.

[2]Welch's t-test provides a measure whether the two distributions are the same, under the assumption that they have normal distributions, but different variances.

[3]On Pentium Xeon 3.6 GHz workstation.

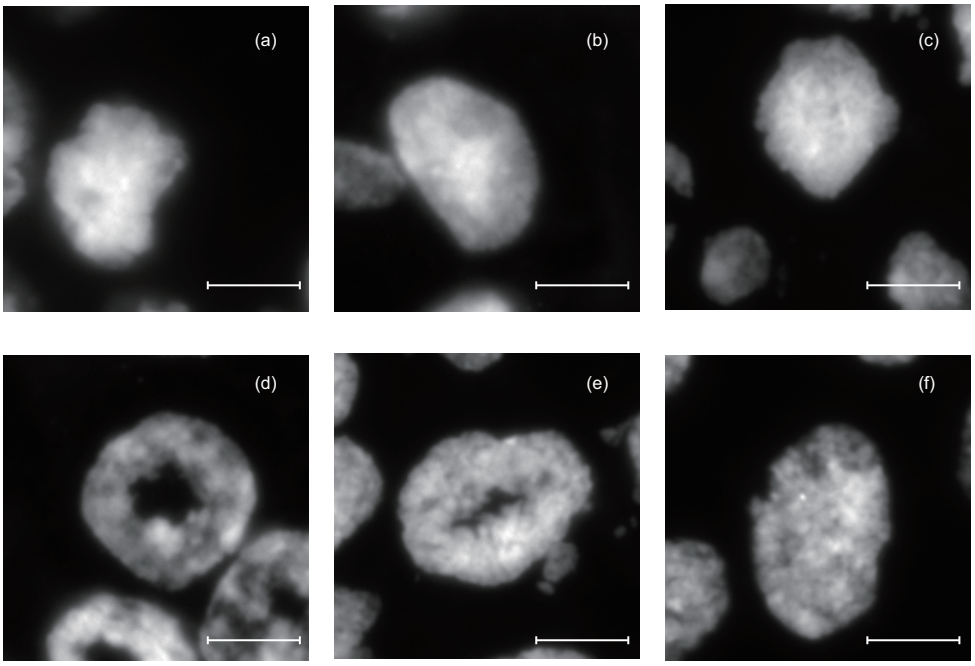

Fig. 6. Examples of cells with low and high contour complexity values. Those cells were collected from the malignant tissues. low value group: (a) 0.007, (b) 0.009, (c) 0.05; high value group: (d) 84.85, (e) 74.98, (f) 58.36. Scale bars, 5μm.

V. CONCLUSIONS AND FUTURE WORK

In this paper, a new approach to chromatin texture characterization, $contour\ complexity$, is proposed and its effectiveness in cancer cell classification is evaluated. This approach is applied to cells collected from benign and malignant breast biopsy tissues. The contour complexity value of malignant cells is significantly greater than that of benign cells.

To achieve automatic feature extraction, we used EM-based segmentation and an effective initial parameter search method for EM. The segmentation method explained in Section III makes it possible to perform the feature extraction in a fully automatic way and applicable to the tissue images stained by other methods, such as hematoxylin-eosine staining and immunostaining. However, a more sophisticated nuclear segmentation method will be needed for overlapping nuclei.

The proposed method appears to grasp different aspects of chromatin texture than fractal analysis, which is widely used in the field of pathology. This means it is more effective when both features are used as a feature vector for machine learning methods. This will be our future work.

It should be emphasized that the results are obtained using paraffin-wax embedded biopsies on slides (i.e., neither frozen sections nor cultured cells), which is actually used for pathological diagnosis. This fact implies that our results are directly connected to real-life cancer diagnosis.

VI. ACKNOWLEDGMENT

The authors gratefully acknowledge Dr. Kenji Okajima, Mr. Atsushi Marugame, Ms. Yoshiko Yamashita and Ms. Maki Ogura for their helpful advice and discussions during the course of this work. T. Kiyuna would also like to thank Dr. Hans Peter Graf of NEC Laboratories America for his insightful comments and suggestions.

REFERENCES

[1] S. Pelengaris and M. Khan (eds.). *The Molecular Biology of Cancer*. Blackwell Publishing, Oxford, UK, 2006.

[2] D. Zink, A. H. Fischer, and J. A. Nickerson. Nuclear structure in cancer cells. *Nature reviews cancer*, 4:677–687, 2004.

[3] R. M. Haralick, K. Shanmugam, and I. Dinstein. Textural features for image classification. *IEEE Transaction on Systems, man and cybernetics*, 3:610–621, 1973.

[4] I. T. Young, P. W. Verbeek, and B. H. Mayall. Characterization of chromatin distribution in cell nuclei. *Cytometry*, 7:467–474, 1996.

[5] K. Yogesan, T. Jørgensen, F. Albregtsen, K. J. Tveter, and H. E. Danielsen. Entropy-based texture analysis of chromatin structure in advanced prostate cancer. *Cytometry*, 24(3):268–276, 1996.

[6] A. J. Einstein, H-S Wu, and J. Gil. Self-affinity and lacunarity of chromatin texture in benign and malignant breast epithelial cell nuclei. *Phys. Rev. Lett.*, 80(2):397–400, 1998.

[7] S. Majumdar, R. S. Weinstein, and R. R. Prasad. Application of fractal geometry techniques to the study of trabecular bone. *Med. Phys.*, 20(6):1611–1619, 1993.

[8] D. Pereira, C. Zambrano, and M. Martin-Landrove. Evaluation of malignancy in tumors of the central nervous systemusing fractal dimension. *Proceedings of the 22nd Annual International Conference of the IEEE*, 3:1775–1778, 2000.

[9] S. S. Cross, J. P. Bury, and P. B. Silcocks. Fractals in pathology. *J. of Pathol.*, 182:1–8, 1997.

[10] S. S. Cross, J. P. Bury, and P. B. Silcocks. Fractal geometric analysis of colorectal polyps. *J. of Pathol.*, 172:317–323, 1994.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc.*, 30:205–248, 1977.

[12] T. Kiyuna, K. Kamijo, T. Yamazaki, N. Moriyama, and R. Sekiguchi. Automated reconstruction of a three-dimensional brain model from magnetic resonance images. *NeuroImage*, 13(6):S173, 2001.

[13] M. Kuroda, H. Tanabe, K. Yoshida, K. Oikawa, A. Saito, T. Kiyuna, M. Hiroshi, and K. Mukai. Alteration of chromosome positioning during adipocyte differentiation. *J. Cell. Sci.*, 117:5897–5903, 2004.