ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Biomedical Informatics 35 (2002) 92–98

Journal of Biomedical Informatics

www.academicpress.com

Cancerous nuclei detection on digitized pathological lung color images

Mohamed Sammouda,[a,*] Rachid Sammouda,[b] Noboru Niki,[a] Naohito Yamaguchi,[c] and Noriyuki Moriyama[d]

[a] *Department of Optical Science and Technology, University of Tokushima, Tokushima, Japan*
[b] *Department of Computer Science, University of Sharjah, United Arab Emirates*
[c] *Division of Cancer Information & Epidemiology, National Cancer Center, Tokyo, Japan*
[d] *Department of Diagnostic Radiology, National Cancer Center, Tokyo, Japan*

Received 9 January 2002

Abstract

In this paper, we propose a methodology (in the form of a software package) for automatic extraction of the cancerous nuclei in lung pathological color images. We first segment the images using an unsupervised Hopfield artificial neural network classifier and we label the segmented image based on chromaticity features and histogram analysis of the RGB color space components of the raw image. Then, we fill the holes inside the extracted nuclei regions based on the maximum drawable circle algorithm. All corrected nuclei regions are then classified into normal and cancerous using diagnostic rules formulated with respect to the rules used by experimented pathologist. The proposed method provides quantitative results in diagnosing a lung pathological image set of 16 cases that are comparable to an expert's diagnosis.

Keywords: Lung pathological color image; Segmentation; Artificial neural networks; Chromaticity; Nuclei extraction

1. Introduction

Lung cancer is becoming a serious disease causing a huge number of deaths in Japan [1,2]. To improve the recovery rate for lung cancer, detection and treatment at early stage of cancer growth is necessary. Several techniques are being used for scanning the lung area, such as X-ray Helical CT technique [3–5]. However, the analysis process of the three-dimensional images obtained by this technique is still suffering from some limitations caused by the scanning process. A definitive diagnosis is ordinarily concluded by histology from biopsy. A biopsy involves the collection of small tissue samples examined under a microscope. By examining the cells of the collected tissue, a pathologist can determine the cancerous cells, which are recognized by morphological and chromatic features of their corresponding nuclei regions. With the development of new computer technologies and image processing algorithms, exciting new applications, such as automatic lung tissue tumor diagnosis, are possible.

Here, we present an original automatic extraction method of the cancerous nuclei regions included in pathological color images of lung tissue. The technique is based on chromaticity features and morphological constraints, collected during the analysis of the color images, which are provided by the pathological division of National Cancer Center in Tokyo. In Section 2, we describe the segmentation method based on an unsupervised Hopfield Neural Network (HNN). In Section 3, we present the labeling process of the segmented image based on histograms analysis of the RGB color space components of the raw image. Section 4 describes the correction procedure of the extracted nuclei regions. In Section 5, we compute the radius of the extracted and corrected nuclei regions, and we formulate the diagnostic rules. Our conclusion and future work are given in Section 6.

* Corresponding author. Fax: +81-886-56-9433.
E-mail address: mohamed@opt.tokushima-u.ac.jp (M. Sammouda).

1532-0464/02/$ - see front matter © 2002 Elsevier Science (USA). All rights reserved.
PII: S1532-0464(02)00501-4

2. Segmentation method

The segmentation problem of an image of N pixels is formulated in [6] as a partition of the N pixels among M classes, such that the assignment of the pixels minimizes a criterion function. HNN has been used in [6] for the segmentation of cerebral images obtained using Magnetic Resonance Imaging (MRI) technique. Some limitations have been reported due to the stuck of the network in an early local minimum and that because the energy landscape in general has more than one local minimum due to the nonconvex nature of the energy surface. To overcome this problem, we have given in our previous work [7] some contributions to the algorithm presented in [6] making HNN converges to a local minimum closer to the global minimum in a pre-specified time. In a comparison of the modified HNN based segmentation algorithm with those of Boltzmann Machine (BM) and ISODATA, HNN showed a clear advantage in giving crisp segmentation of cerebral MRI images as shown in [7].

Similarly to MRI images, color images can be considered as multidimensional data as each pixel can be represented by its three components in each of the RGB or HSV color spaces. Furthermore, in case of color images more information can be extracted from the raw images based on some chromatic properties and added to the features of each pixel of the image. These features can help for more accurate segmentation and labeling of the images to be considered as a base for lung cancer diagnosis.

The HNN structure consists of $N \times M$ neurons with each row representing a pixel and each column representing a cluster. The network is designed to classify the image of N pixels of P features among M classes, such that the assignment of the pixels minimizes the following criterion function:

$$E = \frac{1}{2}\sum_{k=1}^{N}\sum_{l=1}^{M} R_{kl}^2 V_{kl}^2 + c(t)\sum_{k=1}^{N}\sum_{l=1}^{M} N_{kl}V_{kl}, \tag{1}$$

where R_{kl} is considered as the Mahalanobis distance measure between the kth pixel and the centroid of class l and given by

$$R_{kl} = \|X_k - \overline{X}_l\|_{\Sigma_l^{-1}} = \left(X_k - \overline{X}_l\right)^{\mathrm{T}} \Sigma_l^{-1} \left(X_k - \overline{X}_l\right), \tag{2}$$

where X_k is the P-dimensional feature vector of the kth pixel (here, $P = 3$ with respect to the RGB or HSV color space components), $\overline{X}_l$ is the P-dimensional centroid vector of class l, and Σ_l is the covariance matrix of class l.

V_{kl} is the output of the kth neuron. N_{kl} is a $N \times M$ vector of independent high frequency white noise source used to avoid the network being trapped in early local minimums. The term $c(t)$ is a parameter controlling the magnitude of noise, which is selected in a way that it provides zero as the network reaches the convergence. The minimization is achieved by using HNN and by solving the motion equations satisfying

$$\frac{\partial U_{kl}}{\partial t} = -\mu(t)\frac{\partial E}{\partial V_{kl}}, \tag{3}$$

where U_{kl} is the input of the kth neuron, and $\mu(t)$ is a scalar positive function of time, used to increase the convergence speed of HNN, and is defined as in [7] by

$$\mu(t) = t(T_s - t), \tag{4}$$

where t is the iteration step, and T_s is the pre-specified convergence time of the network which has been found to be 120 iterations [7]. The network classifies the feature space, without teacher, based on the compactness of each cluster, calculated using Mahalanobis distance measure between the kth pixel and the centroid of class l.

2.1. Segmentation algorithm

The segmentation algorithm is as follows:

Step 1. Initialize the input of the neurons to random values.

Step 2. Apply the following input–output relation, establishing the assignment of each pixel to only and only one class

$$\begin{aligned} &V_{km}(t+1) = 1 \quad \text{if } U_{km} = \mathrm{Max}[U_{kl}(t)\ \forall l],\\ &V_{kl}(t+1) = 0 \quad \text{otherwise.} \end{aligned} \tag{5}$$

Step 3. Compute the centroid $\overline{X}_l$ and the covariance matrix Σ_l of each class l as follows:

$$\overline{X}_l = \frac{\sum_{k=1}^{N} X_k V_{kl}}{n_l}, \tag{6}$$

$$\Sigma_l = \frac{\lfloor V_{kl}\left(X_k - \overline{X}_l\right)\left(X_k - \overline{X}_l\right)^{\mathrm{T}} \rfloor}{n_l - 1}, \tag{7}$$

where n_l is the number of pixels in class 1. The covariance matrix is normalized by dividing each of its elements by $\det[\Sigma_l]^{1/P}$.

Step 4. Update the inputs of each neuron by solving the set of differential equations in (2) using Eulers approximation:

$$U_{kl}(t+1) = U_{kl}(t) + \frac{\mathrm{d}U_{kl}}{\mathrm{d}t}. \tag{8}$$

Step 5. If $t < T_s$, repeat from Step 2, else terminate.

2.2. Segmentation results

In our previous work [7], we have shown the effectiveness of HNN classifier in segmenting MRI cerebral gray level images. However, the segmentation result of color images does not depend only on the performance of the segmentation algorithm, but also on the color space representation [8]. To choose a convenient color space representation leading to good segmentation of

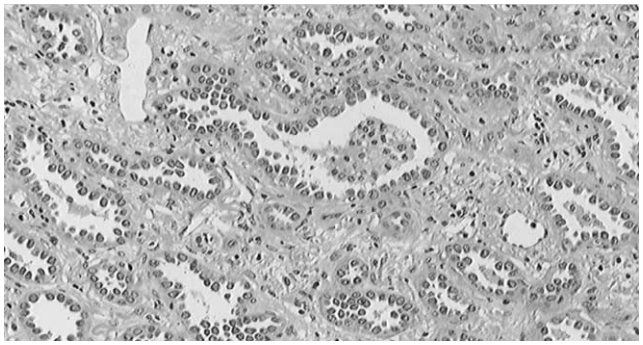

Fig. 1. A sample of pathological lung color image in true color (red, green, blue) (10× magnification).

our data, we segment the same data using HNN with the RGB and HSV representations.

Fig. 1 shows an RGB color image of lung tissue of 960×512 pixels obtained with 10× magnifications. Figs. 2 and 3 show the corresponding segmentation results using HNN classifier with the RGB and HSV representations of the image in Fig. 1, respectively. The number of classes is fixed to three as nuclei, cytoplasm, and others. In each segmentation process HNN converges to the result after 120 iterations in 110 s using color Sun Sparcstation10 with 640 MB memory and running SunOS 5.7.

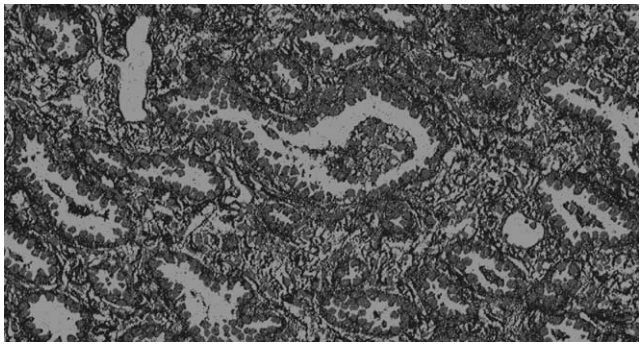

Fig. 2. The segmented result of the lung tissue image shown in Fig. 1 based on RGB color.

Fig. 3. The segmented result of the lung tissue image shown in Fig. 1 based on HSV color.

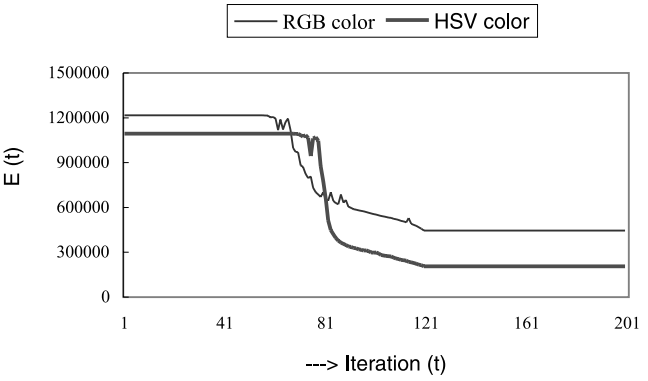

Fig. 4. Energy function curves of the HNN during the optimization process corresponding to the results shown in Figs. 2 and 3.

Visually, the two segmentation results look identical, matching to the raw image we find that in both cases the nuclei are well classified. However, the cytoplasm regions in the RGB based segmentation are classified in two different classes, the cytoplasm and others. This is can be due to the change of the reagent caused by the environment when the photo is taken. This cause is less effective in the HSV representation, as can be seen in the segmentation result in Fig. 3, which matches easily to the raw image in Fig. 1.

Quantitatively, the energy function of HNN, which represents the error committed during the classification process, is smaller in case of using HSV representation as shown in Fig. 4. This confirms the advantage of the use of HSV color representation in segmenting lung tissue color images based on the above described HNN classifier. In this paper, from now on we consider the HSV as our default color space representation of lung tissue color images.

3. Labeling algorithm

Usually, pathologists make their diagnostic decisions by viewing specimen tissue or cells and assessing the various diagnostically important parameters, such as the size, shape, and textures [9]. As the nuclei and the corresponding cytoplasm regions are the most important for the diagnosis process [10,11], we present the following automatic algorithm to extract these regions of interests (ROIs), which is based on the analysis of the histograms of the RGB components of the raw image.

3.1. Histogram analysis

From a set of 16 color images of lung tissue sections of 3 μm from patients with ademocarcinoma, stained with hematoxylin and eosin, we took randomly three images as study cases. To analyze the differences between the three samples, we have superposed the

histograms of their corresponding red, green, and blue components, as shown in Figs. 5A–C, respectively. We realize that only the histograms of their green components conform to each other. This can be explained by the analysis of Fig. 6, which shows the RGB wavelength curves over the spectrum, given in [12] showing that only the red and the blue components of the raw image are affected by the luminosity noise. This special feature motivated us to segment the histogram of the green component into three intervals associated with the main classes of the raw image. Three threshold values, $T_1 = 110$, $T_2 = 210$, and $T_3 = 255$, are found to be as upper limits in the labeling procedure, corresponding to the nuclei, cytoplasm, and others regions, respectively.

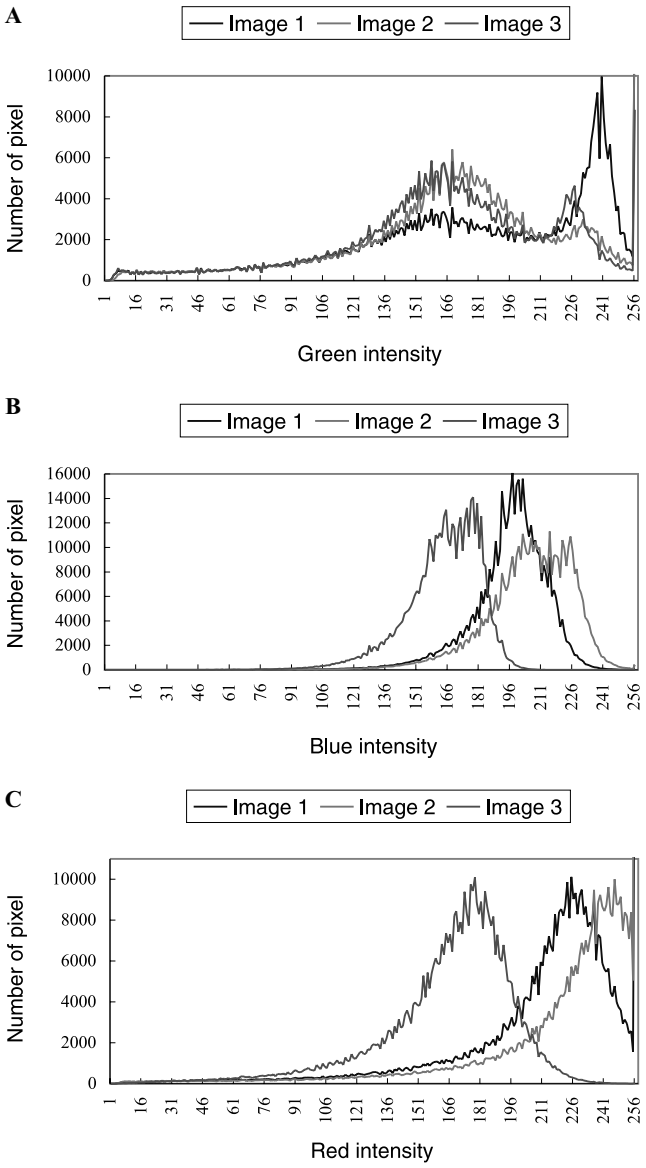

Fig. 5. (A) The green component intensity histograms of the study set. (B) The blue component intensity histograms of the study set. (C) The red component intensity histograms of the study set.

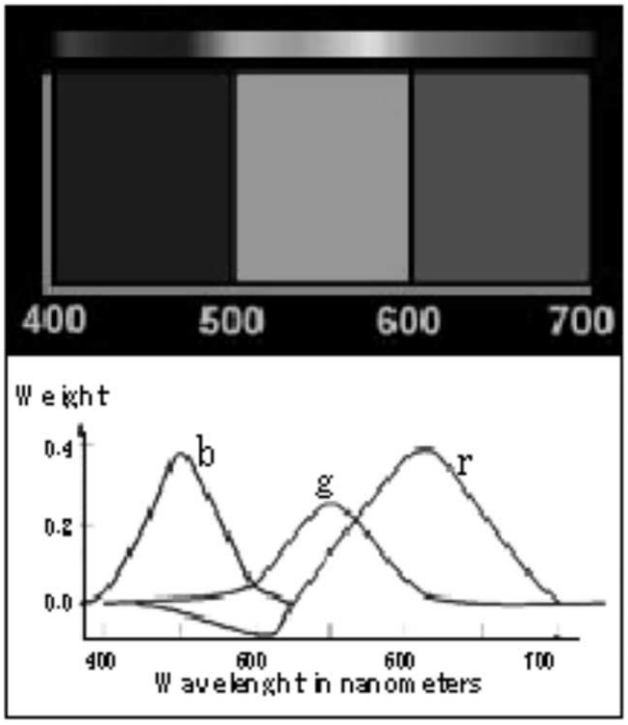

Fig. 6. The RGB wavelength curves over the spectrum of 400–700 nm.

These values have been confirmed after experimental tests during which we were varying them to get good display results in comparison with the raw images.

3.2. Labeling procedure

The segmentation results obtained using HNN with the HSV color space components of the raw images are labeled based on the information collected from the histograms analysis and using the following algorithm:

```
Begin
    For class = 1 To M (M is the number of classes),
    Begin
        Label_Nuclei = 0;
        Label_Cytoplasm = 0;
        Label_Others = 0;
        For Pixel = 1 To NClass (NClass is the
number of pixels in the Class)
        Begin
            if Pixel_Green_Component_Intensity ∈
[0, T1]
              then Label_Nuclei + = 1;
            else
                if Pixel_Green_Component_Inten-
sity ∈]T1, T2]
                  then Label_Cytoplasm + = 1;
                else Label_Others + = 1;
                endif
            endif
        end
```

Label(class) = Maximum(Label _Nuclei, Label _Cytoplasm, Label _Others);
end
end

Figs. 7–9 show the result of the labeling algorithm, corresponding to the nuclei, cytoplasm, and others labels, respectively.

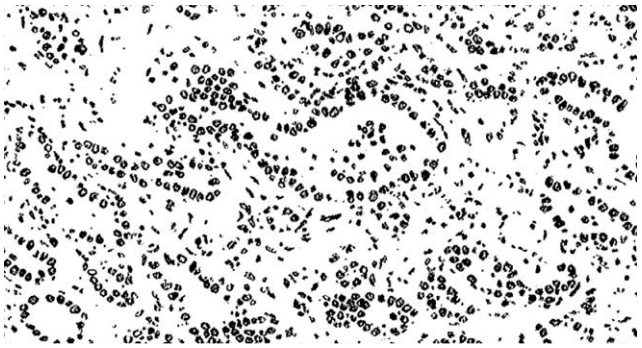

Fig. 7. The automatic extracted nuclei regions (the nuclei regions are shown in black color).

Fig. 8. The automatic extracted cytoplasm regions (the cytoplasm regions are shown in black color).

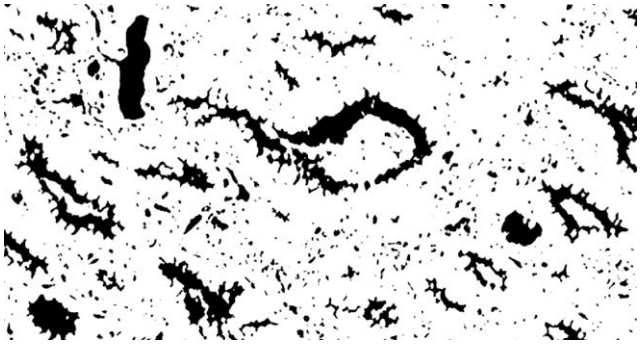

Fig. 9. The automatic extracted regions of the other lung tissue components (the other regions are shown in black color).

4. Correction of the nuclei regions

The result of the extraction and labeling of the nuclei regions shown in Fig. 7 shows some holes inside each nucleus and this can make the diagnostic process more complex. To overcome this problem, we have used the maximum drawable circle algorithm to fill these holes [13]. The algorithm is applied to the image inverse of the extracted nuclei image. Therefore, the background and holes inside nuclei in the nuclei image became objects. Then the maximum radius is calculated for each object, after which all small objects (holes) are removed. Fig. 10 shows a single nucleus before and after correction. Fig. 11 shows the result of the correction procedure applied to the extracted nuclei image of Fig. 7.

5. Cancerous nuclei extraction

When characterizing normal and abnormal cells, a good starting point is to consider the criteria used by human experts in their visual examination of cell samples. Most researchers in automated cytology have concentrated on the classification of isolated single cells [9]. However, even with improved specimen preparation techniques, not all cells are possible to be analyzed as single cells. Indeed, it may be true that the so-called improved techniques destroy the normal relations between cells to such degree that valuable information can be lost [14,15]. Here, we focus on single cell features with respect to the group of its surrounding cells. Furthermore, it has been reported that simple global features describing the size and density of the nuclei can be used as powerful tools for discriminating between normal and cancerous cells [16,17]. This not surprising since it is well known that an abnormal cell contains more DNA than normal one (hyperchromasia), therefore, the cytoplasm of atypical cells are often reduced in size [17]. Here, we follow the idea presented in [15,18], and we compute the nucleus size to be used in extracting cancerous nuclei among the lung tissue.

5.1. Nuclei radius computation

After the correction of the extracted nuclei regions with the algorithm described in the previous section, each pixel value in the corrected image is replaced with the radius of the maximum drawable circle inside the

Fig. 10. The nucleus correction process: (a) object before correction, (b) object after correction, (c) raw nucleus before correction, (d) raw nucleus after correction.

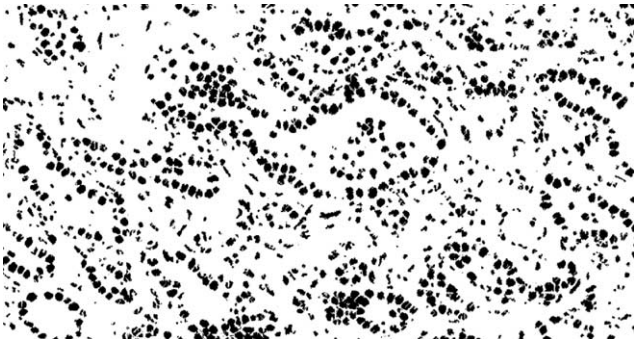

Fig. 11. The nuclei correction result applied to the image shown in Fig. 7.

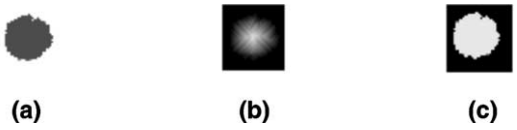

Fig. 12. The maximum drawable circle and radius computing processes: (a) binary nucleus, (b) maximum drawable circle inside nucleus: grayscale, (c) all pixels within nucleus have the maximum value.

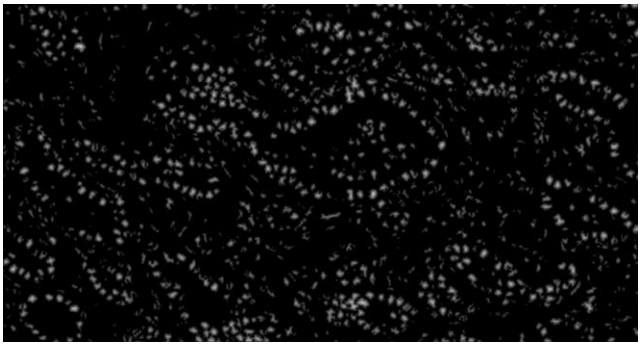

Fig. 13. The maximum drawable circle result applied to the corrected nuclei image shown in Fig. 11: grayscale display.

nucleus region. Later, each pixel belongs to one nucleus, gets the maximum value among all pixels within the same region. By this process, every nucleus has one radius value. Fig. 12 shows the result of this procedure applied to one single nucleus and Fig. 13 shows the result of the same procedure applied to the corrected image of Fig. 11.

5.2. Rules to extract cancerous nuclei

To formulate the diagnostic rules, we have considered a pathologist diagnosis of two random cases taken from our collected data of lung tissue containing normal and abnormal areas. Then we calculated the nuclei sizes (radius) for each sample (one sample with 10 magnifications and another with 20). By matching the pathologist diagnosis to the nuclei sizes, we were able to extract the following diagnostic rules:

Let D be an integer and $D \geqslant 2$ as the radius of one nucleus.

Rule 1: In case of a magnification factor of 10, if $D \notin [2,5[$ then the nucleus is cancerous.

Rule 2: In case of a magnification factor of 20, if $D \notin [2,9[$ then the nucleus is cancerous.

Fig. 14 shows the cancerous nuclei present in the sample of Fig. 1, extracted using the above described rules. Fig. 15 shows the diagnostic result of the same sample given by an expert pathologist. The cancerous regions are marked by black color pen. To evaluate the performance of our algorithm in its first stage, we applied it to 16 different lung cancer (adenocarcinoma) pathology images and the results were compared to that of pathologist results. The evaluation is made by dividing the number of the cancerous nuclei extracted using our system, by the number of the cancerous nuclei extracted visually by the pathologist, and the results are shown in Table 1.

By examining Table 1, we realize that the percentages range between 95% and 97%, which shows that our proposed algorithm gave almost a similar output (the

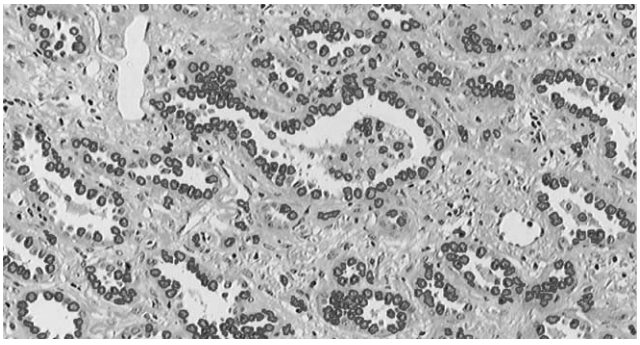

Fig. 14. The cancerous nucleus extraction results using the proposed algorithm applied to the raw image shown in Fig. 1. The cancerous nuclei regions are shown by the dark black color.

Fig. 15. The cancerous nucleus result extracted manually by the pathologist of the image shown in Fig. 1. The cancerous regions are marked by the black color.

Table 1
The system evaluation

Image number	System evaluation (%)
1	97
2	97
3	95
4	97
5	95
6	96
7	96
8	97
9	97
10	97
11	96
12	96
13	95
14	97
15	92
16	92

number of cancerous nuclei) to that given by the pathologist in most study cases, except for cases 15 and 16, and that because image 15 includes many nuclei whose inside materials are invisible, and image 16 includes several clumps consisting of 10–15 nuclei. However, even for these kinds of images, our algorithm performance was 92%. These experimental results indicate that the proposed algorithm is effective for an automatic recognition and extraction of cancerous nuclei in lung tissue color images.

6. Conclusion

In this paper, a new method has been described for an automatic extraction of cancerous nuclei in lung tissue color images. In segmentation process, we have shown that using the HNN classifier with the HSV color space representation of the raw images gives accurate classification of the pixels. The classification results with some information collected based on histogram analysis of the RGB components of the color image are used to make an automatic labeling of the different regions of lung tissue. New diagnostic rules for cancerous nuclei have been formulated based on expert knowledge, showed comparable results to an expert's diagnosis. In our future work, we will test other tissue images from different parts of the human body to generalize the segmentation algorithm and the decision about the number of the classes. Also, we will strength the system with more fine extraction rules in relation with the morphology of the ROIs and enlarge our database with more normal and abnormal lung tissue color images in the sake of constructing a lung cancer diagnosis system.

References

[1] Kaneko M, Eguchi K, Ohmatsu H, Kakinuma R, Naruke T, Suemasu K, Moriyama N. Peripheral lung cancer: screening and detection with low-dose spiral CT versus radiography. Radiology 1996;201:798–802.
[2] National Cancer Center, Tokyo, Japan. Available from: http://www.info.ncc.go.jp/statistics.
[3] Mayo J, Muller NL, Henkelman RM. The double-fissure sign: a motion artifact on thin section CT scans. Radiology 1987;165:50–8.
[4] Terver RD, Conces DL, Godwin JD. Motion artifacts on CT simulate bronchiectasis. Am J Roentgenol 1988;151:1117–9.
[5] Hsiel J. A general approach to the reconstruction of the X-ray helical computed tomography. Med Phys 1996;23(2):221–9.
[6] Amartur SC, Piraino D, Takefuji Y. Optimization neural networks for the segmentation of magnetic resonance images. IEEE Trans Med Imaging 1992;11(2):215–20.
[7] Sammouda R, Niki N, Nishitani H. A comparison of Hopfield Neural Network and Boltzmann machine in segmenting MR images of the brain. IEEE Trans Nucl Sci 1996;43(6).
[8] Liu J, Yee-Hong. Multiresolution color image segmentation. IEEE Trans Pattern Anal Mach Intell 1994;16(7):689–700.
[9] Gil J, Barba J. Morphometry in image analysis for anatomic pathology. In: Marchevsky AM, Bartels PH, editors. Image analysis: a prime for pathologist. NY: Raven; 1994. p. 79–124.
[10] Baak JPA. Manual of quantitative pathology in cancer diagnosis and prognosis. NY: Springer; 1991.
[11] Mayall BH. The monodisperse cell sample: problems and possible solutions. In: Wield GL, Bahr GF, Bartels PH, editors. The automation of uterine cancer cytology. Chicago: Tutorials of cytology; 1976. p. 61–8.
[12] Commission Internationale de l'Eclairage (CIE), Standard on colorimetric observers, CIE S002, 1986.
[13] Hasegawa A, Cullen KJ, Mun SK. Segmentation and analysis of breast cancer pathological images by adaptive-sized hybrid neural network. Proc SPIE, Med Imaging 1996;2710:752–62.
[14] Poulsen RS. Automated prescreening of cervical cytology specimens, Ph.D. Thesis. Montreal: McGill University; 1973.
[15] Spriggs AI, Diamond RA, Mayer EW. A new principal for automated screening of cervical smears. Lancet 1968;1:359.
[16] Crowell RE, Belinsky SA. Genetic changes in lung cancer: potential biomarkers for early detection and prevention. J Lab Clin Med 1997;130:550–7.
[17] Noguchi M, Morikawa A, Kawasaki M. Small ademocarcinoma of the lung: histologic characteristics and prognosis. Cancer 75(12):2844–52.
[18] Diamond RA. Cytology criteria applicable to automatic scanning of cervical smears for malignancy, B.Sc. Thesis. Oxford; 1968.