Unsupervised Segmentation of Overlapped Nuclei Using Bayesian Classification

Chanho Jung, Changick Kim*, *Senior Member, IEEE*, Seoung Wan Chae, and Sukjoong Oh

***Abstract*—In a fully automatic cell extraction process, one of the main issues to overcome is the problem related to extracting overlapped nuclei since such nuclei will often affect the quantitative analysis of cell images. In this paper, we present an unsupervised Bayesian classification scheme for separating overlapped nuclei. The proposed approach first involves applying the distance transform to overlapped nuclei. The topographic surface generated by distance transform is viewed as a mixture of Gaussians in the proposed algorithm. In order to learn the distribution of the topographic surface, the parametric expectation-maximization (EM) algorithm is employed. Cluster validation is performed to determine how many nuclei are overlapped. Our segmentation approach incorporates *a priori* knowledge about the regular shape of clumped nuclei to yield more accurate segmentation results. Experimental results show that the proposed method yields superior segmentation performance, compared to those produced by conventional schemes.**

***Index Terms*—Automatic cell segmentation, cluster validation, Gaussian mixture model, overlapped nuclei segmentation, unsupervised Bayesian classifier.**

I. Introduction

QUANTITATIVE microscopic cell image analysis is often desired in diagnostic and experimental pathologies for objective and consistent evaluation. Interaction-aware methodologies have been the most widely used techniques to assess the cell image [1]–[3]. However, these approaches are time consuming and often infeasible [1], [3]. Moreover, the diagnostic information produced by human interaction is subjective and irreproducible [2], [4]. In recent years, a number of fully automatic cell image analysis methods have been introduced to address the problems [5]–[7].

Cell image segmentation is one of the main functional components in a fully automated cell image analysis [2], [4]. In the literature, numerous cell image segmentation techniques have been reported [1], [8]. However, separating overlapped or clumped nuclei still remains a challenging task in the field of cell segmentation [1]–[3], [9]–[11]. In addition, the performance of cell segmentation is extremely dependent on the ability of segmenting overlapped nuclei. A variety of schemes using curvature information have been investigated to separate the overlapped nuclei [12]–[14]. However, most of these methods are highly dependent on the concavity, and thus it is difficult to obtain robust segmentation results when the concave points are not detected correctly [12]–[14]. The watershed algorithm is the most popular segmentation scheme to handle the problems [2], [3]. Since a regional minimum corresponds to an object, the watershed transform usually leads to oversegmentation. Region merging and marker-controlled watershed techniques have been introduced to prevent the oversegmentation [1], [3]. The region merging approaches tend to rely on shape and size of nuclei, whereas the marker-controlled watershed has an ability to reduce such problems by using marker extraction. Furthermore, it only works well when the extracted markers represent true objects correctly [1], [3]. Complement of distance map has been used to avoid the oversegmentation. However, the nuclei cannot be separated when they are severely overlapped or clumped [2]. Mathematical morphology could be employed to eliminate spurious markers, but the segmentation results are sensitive to incorporated structuring elements and thresholds [3], [15].

In this paper, to alleviate the aforementioned problems, we propose an unsupervised Bayesian classification methodology to separate overlapped or clumped nuclei. Distance transform provides a measure for separating clustered objects. The distance image can be interpreted as a topographic surface. The topographic surface generated by distance transform is viewed as a mixture of Gaussians in this paper. Given the type of distribution, we employ the parametric expectation-maximization (EM) technique to estimate parameters of the Gaussian mixture model (GMM). In other words, we formulate the separation of overlapped nuclei as a cluster analysis problem. In order to investigate the cluster, the unsupervised Bayesian classifier is adopted. Then, cluster validity index is employed to estimate the number of nuclei overlapped. Since the Gaussian distributions are overlapped, *separation* and *compactness* of the mixture are used to validate the cluster.

Manuscript received June 23, 2009; revised March 24, 2010; accepted June 26, 2010. Date of publication July 23, 2010; date of current version November 17, 2010. This work was supported by the Ministry of Knowledge Economy (MKE), Korea, under the Information Technology Research Center (ITRC) support program supervised by the National IT Industry Promotion Agency (NIPA) [NIPA-2010-(C1090-1011-0003)]. *Asterisk indicates corresponding author.*

C. Jung is with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 305-732, Korea (e-mail: peterjung@kaist.ac.kr).

*C. Kim is with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 305-732, Korea (e-mail: cikim@ee.kaist.ac.kr).

S. W. Chae is with the Department of Pathology, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, Seoul 440-746, Korea (e-mail: swan.chae@samsung.com).

S. Oh is with the Department of Internal Medicine, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, Seoul 440-746, Korea (e-mail: sukjoong.oh@samsung.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBME.2010.2060486

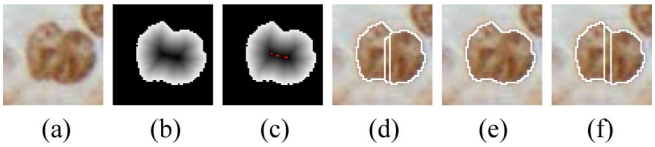

Fig. 1. Separation results of several marker-controlled watershed methods: (a) original image, (b) complement of distance image, (c) regional minima, (d) classical watershed, (e) condition erosion [3], and (f) adaptive *H*-minima transform [1].

Typically, incorporating *a priori* knowledge of the overlapped nuclei improves the segmentation performance [16], [17]. In our study, the geometric property of the separation line and the shape of general nuclei are taken into account within the framework. To obtain the hyperplane, which maximizes the separability among clusters, linear discriminant analysis (LDA) is employed. By using partial contours from the nonoverlapped regions, contour reconstruction in the overlapped regions can be performed. A constrained ellipse fitting is introduced to reconstruct the contours of the occluded regions.

The rest of the paper is organized as follows: In Section II, the proposed method for separating overlapped nuclei is presented. Experimental results are shown in Section III, followed by concluding remarks in Section IV.

II. Proposed Method

A. Motivation

As described in the previous section, the marker-controlled watershed technique is known as one of the most robust solutions for separating overlapped or clumped nuclei [2], [3], [9], [10]. However, the spurious markers are unavoidable when the nuclei are severely overlapped. *H*-minima transform and morphological reconstruction techniques can be used to alleviate the problems [2], [18]. Fig. 1 shows separation results of several marker-controlled watershed methods for severely overlapped nuclei. As we can see in Fig. 1(d) and (f), the oversegmentation is observed due to the spurious marker. Moreover, undersegmentation is occurred by incorrect marker extraction, as shown in Fig. 1(e).

The aim of this paper is to present a framework for separating overlapped or clumped nuclei minimizing oversegmentation and undersegmentation. Opening by reconstruction and closing by reconstruction are sequentially applied to remove unwanted noise without changing the shape of overlapped nuclei. To perform the distance transform, adaptive thresholding is applied to the sequentially filtered image [19]. Fig. 2(b) and (c) shows the result of sequential filtering and adaptive thresholding for overlapped nuclei, respectively. In our study, we observe that the distance image of overlapped nuclei shows a Gaussian mixture distribution, as shown in Fig. 2(e). The distance image, as shown in Fig. 2(d), is obtained by measuring the minimum distance between each pixel and the pixel that does not belong to the overlapped nuclei. Since we know the distribution structure of overlapped nuclei, the distribution can be learned in an unsupervised manner.

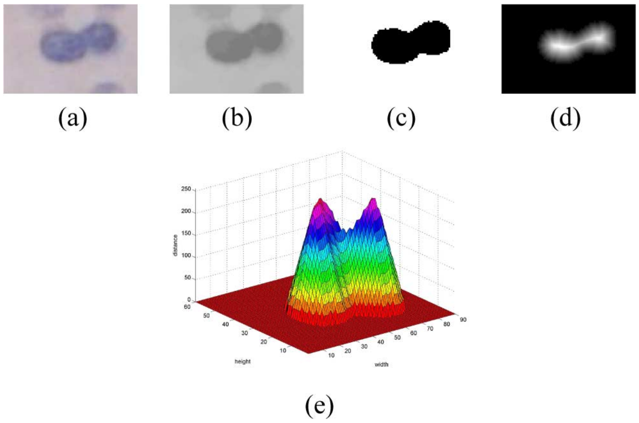

Fig. 2. Distance image of overlapped nuclei: (a) original image, (b) sequentially filtered image, (c) binary image (d) distance image, and (e) topographic surface.

B. Unsupervised Bayesian Classification

In order to estimate the GMM, the parametric EM algorithm is employed. Let $\mathbf{x} = [x\ y]^t$ be an observed sample, which is a pixel position that has a distance value larger than zero, extracted from the topographic surface obtained by distance transform for overlapped nuclei. Let $L(\mathbf{x}_i)$ be the distance to the nearest background pixel from the ith observed sample $\mathbf{x}_i$. In order to relate the topographic surface with the mixture of Gaussian distributions, we interpret that $\mathbf{x}_i$ is observed $L(\mathbf{x}_i)$ times. Then, the number of unlabeled-independent observed samples in our GMM can be expressed as

$$n = \sum_{i=1}^{l} L(\mathbf{x}_i) \tag{1}$$

where l represents the number of pixels belonging to the overlapped nuclei. The Gaussian mixture density for the observed samples is expressed by

$$p(\mathbf{x}|\boldsymbol{\theta}) = \sum_{j=1}^{m} p(\mathbf{x}|\omega_j, \boldsymbol{\theta}_j) P(\omega_j) \tag{2}$$

where $p(\mathbf{x}|\omega_j, \boldsymbol{\theta}_j)$ and $P(\omega_j)$ represent the class-conditional probability density and the prior probability, respectively, and ω_j and m denote the jth nucleus and the number of nuclei, respectively. In (2), $\boldsymbol{\theta}$ represents the unknown parameter vector including the mean vectors and the covariance matrices. The class-conditional probability density is described as

$$p(\mathbf{x}|\omega_j, \boldsymbol{\theta}_j) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_j|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_j)^t \boldsymbol{\Sigma}_j^{-1} (\mathbf{x}-\boldsymbol{\mu}_j)\right] \tag{3}$$

where $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ denote the mean vector and the covariance matrix, respectively, and d represents the dimensionality of observed sample. In our study, $\boldsymbol{\mu}_j$, $\boldsymbol{\Sigma}_j$, and $P(\omega_j)$ are all unknown, and no constraints are placed on the covariance matrix.

We evaluate the likelihood of observed samples, which is defined as shown in (4), to obtain the unknown parameters

$$p(X|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta}) \tag{4}$$

where $p(X|\boldsymbol{\theta})$ denotes the likelihood of $\boldsymbol{\theta}$ with respect to the observed samples. The unknown parameters, which maximize the likelihood, can be estimated by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ln p(X|\boldsymbol{\theta}) \tag{5}$$

where $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ and $\ln p(X|\boldsymbol{\theta})$ represent the sample-set and the log-likelihood function, respectively.

The parametric EM algorithm estimates the posterior probabilities and the unknown parameters iteratively by maximizing the log-likelihood function. On each iteration of the algorithm, two processes are involved: 1) the E-step and 2) the M-step.

Typically, the clustering performance of EM algorithm is highly dependent on the initialization of mixture parameters. Moreover, it is extraordinarily important to find the proper number of clusters in the EM. In this paper, the EM algorithm starts with an initial guess using a set of seeds. Let $\mathbf{s}$ denote a seed. We build a set $R = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_m\}$ of m seeds, each of which corresponds to each nucleus. We will show how to determine the optimal number of seeds, $\hat{m}$, in Section II-C. The regional maxima of the distance map are employed to incorporate the structure of overlapped or clumped nuclei. In order to estimate $\hat{\boldsymbol{\theta}}^0$, X is divided into m subsets via nearest-neighbor algorithm as follows:

$$X_i = \{\mathbf{x} \mid i = \arg\min_k \|\mathbf{x} - \mathbf{s}_k\|^2, \quad \mathbf{x} \in X, 1 \leq k \leq m\}. \tag{6}$$

The mixing proportions are assumed to be uniform.

C. Cluster Validation

When the nuclei are severely clumped, one-to-one correspondence between the regional maxima and the nuclei cannot be established. Therefore, the number of nuclei overlapped cannot be estimated accurately. In the proposed method, the cluster validation is used to evaluate how many nuclei are overlapped. The cluster validity index, which is defined in (7), consists of the *separation* and *compactness* of nuclei [20]

$$V(m) = \frac{\varepsilon(m)}{\varphi(m)} \tag{7}$$

where $\varepsilon(m)$ and $\varphi(m)$ represent the *separation* and *compactness*, respectively. The *separation* is given by

$$\varepsilon(m) = \text{trace}(\mathbf{S}) \tag{8}$$

where $\mathbf{S}$ denotes the weighted between-cluster scatter matrix and is defined as

$$\mathbf{S} = \sum_{i=1}^{m} \sum_{k=1}^{n} p_{ki}^{r} (\mathbf{m}_i - \bar{\mathbf{m}})(\mathbf{m}_i - \bar{\mathbf{m}})^t \tag{9}$$

where p_{ki} represents the posterior probability. In (9), $\mathbf{m}_i$ and $\bar{\mathbf{m}}$ denote the center of ith cluster and the center of overall observed samples, respectively. The *compactness* is given as

$$\varphi(m) = \sum_{i=1}^{m} \text{trace}(\mathbf{C}_i) \tag{10}$$

where $\mathbf{C}_i$ denotes the normalized covariance matrix. $\mathbf{C}_i$ is defined as

$$\mathbf{C}_i = \frac{\sum_{k=1}^{n} p_{ki}^{r} (\mathbf{x}_k - \mathbf{m}_i)(\mathbf{x}_k - \mathbf{m}_i)^t}{\sum_{k=1}^{n} p_{ki}^{r}}. \tag{11}$$

In our study, we chose $r = 1$. To suppress spurious regional maxima, the *H*-maxima transform [18] is employed in building a set of seeds, as shown in (12)

$$H_{\lambda}(\mathbf{I}) = \Phi_{\mathbf{I}}^{\delta}(\mathbf{I} - \lambda) \tag{12}$$

where $\mathbf{I}$ and λ represent the distance map and given depth, respectively, and Φ and δ represent the reconstruction and dilation operators, respectively. The optimal number of nuclei can be obtained by maximizing the cluster validity index as follows:

$$\hat{m} = \arg\max_{m} \frac{\epsilon(m)}{\varphi(m)}. \tag{13}$$

The cluster validity index is estimated if and only if the number of regional maxima for $H_{\lambda}(\mathbf{I})$ equals to m. The overall cluster validation can be summarized as follows:

```
1:  Cluster Validation Procedure:
2:  $m \leftarrow$ # regional maxima
3:  $V_{\max} \leftarrow -\infty$
4:  $\lambda \leftarrow 0$
5:  while $m > 1$ do
6:      repeat
7:          $\lambda \leftarrow \lambda + 1$
8:          evaluate $H_{\lambda}(\mathbf{I})$
9:      until # regional maxima for $H_{\lambda}(\mathbf{I}) \geq m$
10:     if # regional maxima for $H_{\lambda}(\mathbf{I}) = m$ then
11:         construct $R$ from $H_{\lambda}(\mathbf{I})$
12:         perform EM algorithm
13:         estimate $V(m)$
14:         if $V(m) > V_{\max}$ then
15:             $V_{\max} = V(m)$, $\hat{m} = m$, $\hat{R} = R$
16:         end if
17:     end if
18:     $m \leftarrow m - 1$
19: end while
```

In the beginning, the initial number of regional maxima is obtained from the distance image of overlapped nuclei. The cluster validation test is iteratively performed until m equals to two starting from the initial quite large number. Note that line 11 to 16 is skipped when there is no λ satisfying the constraint in line 10. In this case, *m* is decreased by line 18 and then λ satisfying line 10 is recomputed for the decreased m.

D. Separation Line Decision

The watershed-based techniques usually produce jagged contours in the overlapped region. Outer distance transform has been introduced to avoid the jaggedness [1]. Therefore, the true

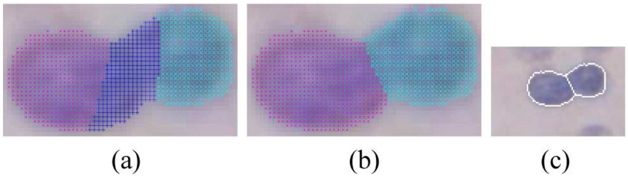

Fig. 3. Estimation of the optimal number of nuclei for overlapped nuclei: (a) clustering result when $m = 3$ ($V = 10.8 \times 10^4$), (b) clustering result when $m = 2$ ($V = 13.2 \times 10^4$), and (c) derived separation line when $m = 2$.

separation line can be considered as a hyperplane dividing the cluster. In order to maximize the separability of overlapped or clumped nuclei, the LDA is adopted for a pair of nuclei touched. We project the observed samples in a direction for which the projected observed samples are well separated. The projection is performed as follows:

$$y = \mathbf{w}^t \mathbf{x} \tag{14}$$

where $\mathbf{w}$ denotes the orientation selected. The orientation is estimated by minimizing the following criterion function:

$$J(\mathbf{w}) = \frac{\mathbf{w}^t \mathbf{S}_B \mathbf{w}}{\mathbf{w}^t \mathbf{S}_W \mathbf{w}} \tag{15}$$

where $\mathbf{S}_B$ and $\mathbf{S}_W$ represent the between-class scatter matrix and within-class scatter matrix, respectively.

To evaluate the separation line, a minimum-error-rate classification scheme is applied to the projected observed samples [21]. Hence, the projected observed samples are classified based on Bayesian decision theory. The discriminant function is defined as

$$g_i(y) = \ln p(y|\omega_i) + \ln P(\omega_i). \tag{16}$$

The estimated separation line should be perpendicular to $\mathbf{w}$. Fig. 3 shows the clustering result and the derived separation line. As shown in Fig. 3(a) and (b), the clustering validity index is maximized when m equals two. Moreover, the proposed method produces the separation line without the jaggedness, as shown in Fig. 3(c).

E. Occluded Contour Reconstruction

Although the contours without the jaggedness can be obtained by using the LDA, the nuclei generally have ellipse-like shaped contours [17]. Since we know the partial contours except for contours in the overlapped region, occluded contours in the overlapped region can be reconstructed. The constrained ellipse fitting technique is employed for reconstruction. Let (p, q) be a point on the partial contour. An ellipse can be described as an implicit second-order polynomial

$$F(p, q) = ap^2 + bpq + cq^2 + dp + eq + f = 0 \tag{17}$$

where a, b, c, d, e, and f represent the ellipse coefficients. $F(p, q)$ can be written to the vector form, which is shown in (18), by defining $\mathbf{a} = [a\, b\, c\, d\, e\, f]^t$ and $\mathbf{p} = [p^2\; pq\, q^2\; p\, q\, 1]^t$.

$$F_{\mathbf{a}}(\mathbf{p}) = \mathbf{a} \cdot \mathbf{p} = 0. \tag{18}$$

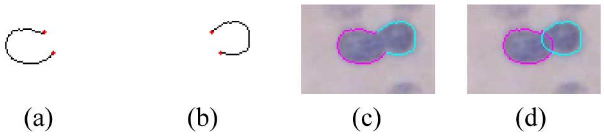

Fig. 4. Partial contours and reconstructed occluded contours: (a) $\mathbf{p}_{os_1}$ and $\mathbf{p}_{oe_1}$, (b) $\mathbf{p}_{os_2}$ and $\mathbf{p}_{oe_2}$, (c) partial contours, and (d) reconstructed occluded contours.

We can find the optimal coefficients $\hat{\mathbf{a}}$ using

$$\hat{\mathbf{a}} = \arg\min_{\mathbf{a}} \|\mathbf{D}\mathbf{a}\|^2 \text{ subject to } \mathbf{a}^t\mathbf{C}\mathbf{a} = 1 \tag{19}$$

where $\mathbf{D}$ and $\mathbf{C}$ denote the design matrix and 6×6 constraint matrix, respectively. The matrices are defined as follows:

$$\mathbf{D} = \begin{bmatrix} p_1^2 & p_1 q_1 & q_1^2 & p_1 & q_1 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_c^2 & p_c q_c & q_c^2 & p_c & q_c & 1 \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{20}$$

where c represents the number of points on the partial contour. $\mathbf{C}$ guarantees that the final solution is an ellipse. The Lagrange multiplier method is used to obtain $\hat{\mathbf{a}}$ [22].

We define a new constraint for occluded contour reconstruction as

$$\sum_i F_{\mathbf{a}}(\mathbf{p}_{os_i}) = 0$$
$$\sum_i F_{\mathbf{a}}(\mathbf{p}_{oe_i}) = 0 \tag{21}$$

where $\mathbf{p}_{os_i}$ and $\mathbf{p}_{oe_i}$ represent the start point and end point of occluded contour, respectively. The constraint matrix in (19) is modified to impose the constraint in (21) as follows:

$$(4ac - b^2) + \sum_i F_{\mathbf{a}}^2(\mathbf{p}_{os_i}) + \sum_i F_{\mathbf{a}}^2(\mathbf{p}_{oei}) = 1. \tag{22}$$

Since the squared algebraic distance is employed in (22), the symmetric property of constraint matrix is preserved. By using (19) and (22), the occluded contours in the overlapped region can be obtained. Specifically, it is ensured that the reconstructed contours pass through $\mathbf{p}_{os_i}$ and $\mathbf{p}_{oe_i}$ by (21) and (22), since $\sum_i F_{\mathbf{a}}^2(\mathbf{p}_{os_i})$ and $\sum_i F_{\mathbf{a}}^2(\mathbf{p}_{oei})$ are zero. Fig. 4 shows the partial contours and the reconstructed occluded contours.

III. Experimental Results

The effectiveness of our method is assessed on real microscopic cell images. A number of specimens containing mammary invasive ductal carcinomas and cervical cells taken by a light microscope with 40x objective are demonstrated. The specimens of mammary invasive ductal carcinomas are stained

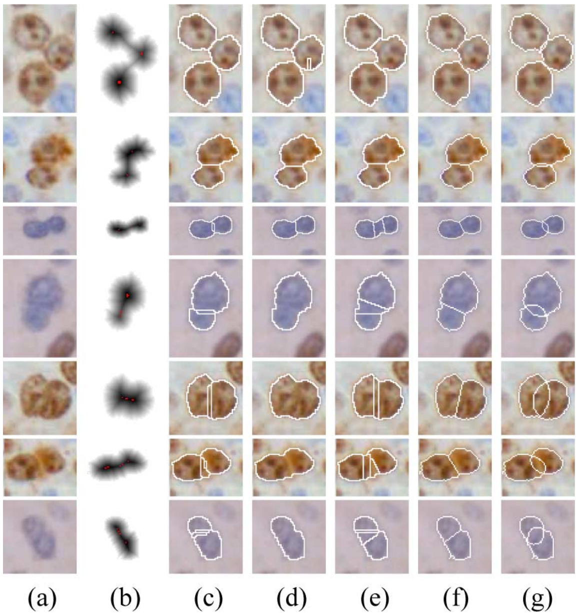

Fig. 5. Comparison of the proposed method with watershed-based separation methods: (a) original image, (b) regional minima, (c) classical watershed, (d) condition erosion [3], (e) adaptive *H*-minima transform [1], (f) proposed method, and (g) reconstructed occluded contours by the proposed method.

by immunohistochemistry for p53 protein and estrogen receptor. Papanicolaou technique is used to stain the specimens of cervical cells. We compare our results with those of various watershed-based segmentation techniques, such as the classical watershed, condition erosion, and adaptive H-minima transform [1], [3]. The results are further compared with ones from parametric fitting algorithm and concavity-based separation methods [13], [14], [17].

Several segmentation results by the proposed separation scheme and the conventional three types of watershed-based segmentation methods are given in Fig. 5. The first and second rows show the separation results for two types of slightly touched nuclei. Since overlapping is not severe, the nuclei are isolated reasonably by the most of separation techniques, but the oversegmentation is observed in (d) of the first row due to the incorrect marker extraction by the condition erosion. The third to seventh rows show that the spurious minima are produced in the overlapped region. Since the minima cannot represent the cells correctly, the classical watershed and adaptive *H*-minima transform yield the oversegmentation, whereas the condition erosion results in the undersegmentation except for the third row due to the drastic overlapping, which makes it difficult to extract the marker accurately. On the other hand, the proposed separation scheme not only segments the overlapped nuclei correctly, but also provides the separation lines reducing the jaggedness. Moreover, the proposed contour reconstruction technique produces the true ellipse-like contour of nucleus, which is useful for quantitative assessment.

The proposed method can be straightforwardly extended to an automated cell image segmentation system. To this end, it is essential to classify whether a connected component on the specimen corresponds to a single nucleus or overlapped nuclei. In the literature, several concavity analysis methods have been widely adopted for the categorization [9], [12], [14]. In our study, the convex-hull-based classification scheme [14] is employed. Table I shows the comparison of segmentation performances for 751 overlapped nuclei on 70 microscopic cell images of size 1392×1040, which consist of 63 specimens of cervical cells and 7 specimens of mammary invasive ductal carcinomas. Table II shows the comparison of segmentation performances for 4635 nuclei including aforementioned 751 overlapped nuclei on the datasets. In our experiments, to make fair evaluations, the categorization step is applied to the classical watershed. This is because the classical watershed is highly likely to yield oversegmentation for single nucleus, which, however, is usually correctly segmented by the condition erosion and adaptive *H*-minima transform schemes. As shown in Table II, the proposed system achieves improvements of up to 6.80%, 5.70%, and 3.43% with respect to the classical watershed, condition erosion, and adaptive *H*-minima transform schemes in terms of separation accuracy, respectively. Note that, as shown in Tables I and II, the success of nuclei segmentation is highly dependent on the segmentation accuracy for the overlapped nuclei. Fig. 6(a) shows parts of our specimens. As shown in Fig. 6, our nuclei segmentation system outperforms the conventional watershed-based approaches. Fig. 7 illustrates how many iterations are required for convergence in the parametric EM. In our study, the termination tolerance for the objective function and the maximum number of iterations allowed are set to be 10^{-6} and 300, respectively. In the figure, *x*-axis represents the row indexes of the overlapped nuclei in Fig. 5. As we can see in Fig. 7, the number of iterations is less than 50 when m is the optimal number of nuclei. When m is not optimal, however, the number of iterations is larger than the one by the optimal number of nuclei. This is because the parametric EM is extremely dependent on the number of mixture components.

Fig. 8 shows a segmentation result of the proposed method compared with the parametric fitting algorithm [17]. In the parametric fitting algorithm, the human interaction is required for providing the number of nuclei overlapped to the separation system. First, the fitting is carried out on a gray-level image, as addressed in [17]. Fig. 8(b) shows that the result of fitting cannot represent the nuclei accurately. Then, the parameters are fitted on binary image. Fig. 8(c) and (d) shows the segmentation results with respect to two different initialization conditions. In the figure, it is observed that the parametric fitting algorithm is very sensitive to the initial parameters. Meanwhile, as shown in Fig. 8(e), the proposed algorithm yields the correct segmentation result without user interaction. In Fig. 9, five segmentation results from the proposed scheme are compared with those from the concavity-based algorithms [13], [14] and the parametric fitting technique [17]. In the concavity-based algorithms, the concave points should be extracted accurately. When the nuclei drastically overlapped, however, it is difficult to detect the concave points correctly. The first to third rows show that the undersegmentation is occurred since the concave points are not extracted. The fourth and fifth figures of Fig. 9(c) show that the overlapped nuclei are separated into three distinct nuclei. However, the segmentation results are not coherent with the

TABLE I
COMPARISON OF SEGMENTATION PERFORMANCES FOR OVERLAPPED NUCLEI ON SPECIMENS OF CERVICAL CELLS AND MAMMARY INVASIVE DUCTAL CARCINOMAS

	correctly segmented	over-segmented	under-segmented
classical watershed	51.53%	47.40%	1.07%
condition erosion [3]	60.27%	1.36%	38.37%
adaptive H-minima transform [1]	72.57%	15.78%	11.65%
the proposed method	93.48%	4.92%	1.60%

TABLE II
COMPARISON OF SEGMENTATION PERFORMANCES ON SPECIMENS OF CERVICAL CELLS AND MAMMARY INVASIVE DUCTAL CARCINOMAS

	correctly segmented	over-segmented	under-segmented
classical watershed	90.72%	9.10%	0.18%
condition erosion [3]	91.82%	1.96%	6.22%
adaptive H-minima transform [1]	94.09%	4.03%	1.88%
the proposed method	97.52%	2.22%	0.26%

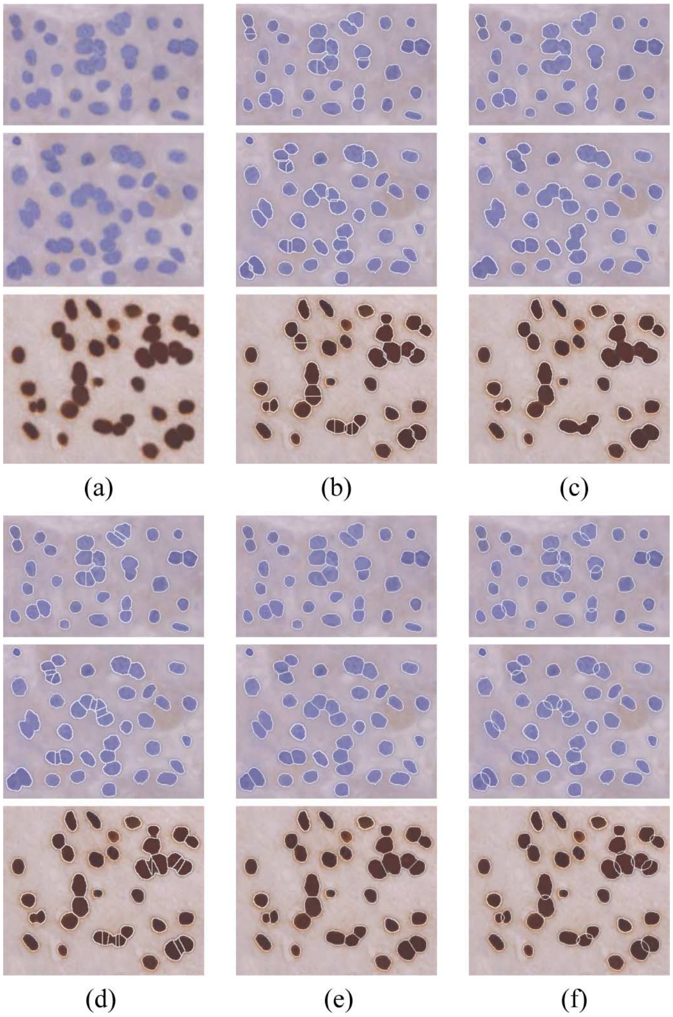

Fig. 6. (a) Original image, (b) classical watershed, (c) condition erosion [3], (d) adaptive *H*-minima transform [1], (e) proposed method, and (f) reconstructed occluded contours by the proposed method.

corresponding binary images. And the occluded contour cannot be reconstructed since the model-based method [13] uses a protocol based on the distance transform to partition squeezed hyperquadrics. Fig. 9(d) and (e) shows the segmentation results obtained from the parametric fitting algorithm and the proposed method. As shown in the figures, both methods segment the

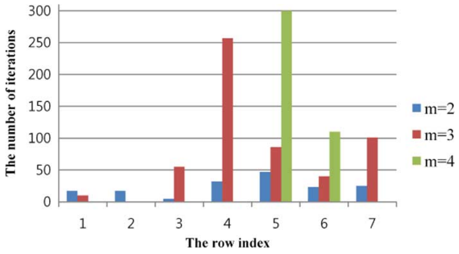

Fig. 7. Number of iterations to convergence in the parametric EM.

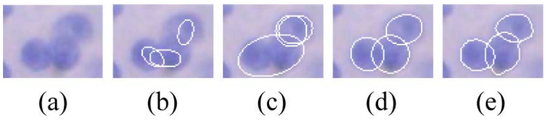

Fig. 8. Comparison of the proposed method with parametric fitting algorithm: (a) original image, (b) fitting result on gray-level image [17], (c) the first fitting result on binary image, (d) the second fitting result on binary image, and (e) reconstructed occluded contours by the proposed method.

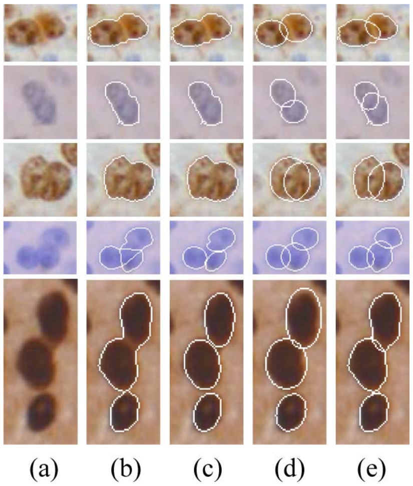

Fig. 9. Comparison of the proposed method with concavity-based methods and the parametric fitting algorithm: (a) original image, (b) rule-based approach [14], (c) model-based segmentation [13], (d) parametric fitting algorithm [17], and (e) reconstructed occluded contours by the proposed method.

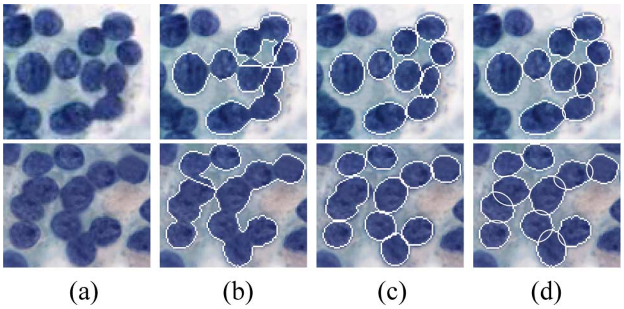

Fig. 10. Comparison of the proposed method with concavity-based methods for cell images with larger number of overlapped nuclei: (a) original image, (b) rule-based approach [14], (c) model-based segmentation [13], and (d) reconstructed occluded contours by the proposed method.

clumped nuclei correctly. In contrast to the proposed method, however, parametric fitting does not represent the nuclei boundary correctly since the whole segments of contours are estimated by the parametric fitting method. In order to evaluate the segmentation performance quantitatively, a segmentation distortion evaluation metric [23], defined as follows:

$$d(\mathbf{N}, \mathbf{N}_G) = \frac{\sum_{\mathbf{x}} \mathbf{N}(\mathbf{x}) \oplus \mathbf{N}_G(\mathbf{x})}{\sum_{\mathbf{x}} \mathbf{N}_G(\mathbf{x})} \tag{23}$$

is employed, where $\mathbf{N}$ and $\mathbf{N}_G$ represent the nuclei segmentation result and the ground truth, respectively, and $\oplus$ represents the Boolean exclusive-OR operator. The ground truth is obtained by manual segmentation. The segmentation distortions of our method for the segmentation results in Fig. 9 are compared with those of the rule-based [14] and model-based [13] methods, which are also unsupervised approaches. The average distortion of the proposed method was 0.14, whereas those of the rule-based [14] and model-based [13] approaches were 0.65 and 0.51, respectively. Fig. 10 shows the comparison of separation results when the cluster consists of a large number of nuclei. As shown in Fig. 10, the number of nuclei overlapped has little influence on the segmentation performance of proposed method, even though the complexity of EM increases with the number of nuclei overlapped.

IV. Conclusion

In this paper, we proposed a fully automated overlapped nuclei separation scheme, using unsupervised Bayesian classifier. A GMM was introduced to investigate distance image of the overlapped nuclei. Since the form of distribution was known, parametric EM algorithm was employed to learn the GMM. In order to evaluate the optimal number of nuclei, cluster validation was performed based on the extracted regional maxima. *A priori* knowledge for the overlapped nuclei was incorporated to obtain separation line without jaggedness, as well as to reconstruct occluded contours in overlapped region. Experimental results presented in Section III prove that the proposed separation methodology segments better than the conventional ones including a classical watershed, a condition erosion, an adaptive *H*-minima transform, a parametric fitting algorithm, a rule-based approach, and a model-based segmentation. In addition, it turns out that the proposed occluded contour reconstruction scheme can produce ellipse-like shaped contour close to original one, which is needed for quantitative analysis of the microscopic cell images.

Although the proposed method outperforms the conventional approaches with respect to segmentation accuracy, it should also be noted that the computational complexity is still high due to employing the parametric EM for unsupervised cluster analysis. Since the cluster analysis approach was proved to be a powerful technique for nuclei segmentation in this paper, however, our future work is to improve processing time for high-throughput analysis of cell images. Currently, our implementation on Intel Core2 Duo 3 GHz PC takes 2.26 s/image, on average. On the other hand, the proposed algorithm can be parallelized on multiple CPUs for significant speedup.

Acknowledgment

The authors would like to thank the Department of Pathology, Kangbuk Samsung Hospital, Seoul, Korea, for providing the dataset of microscopic cell images.

References

[1] J. Cheng and J. C. Rajapakse, “Segmentation of clustered nuclei with shape markers and marking function,” *IEEE Trans. Biomed. Eng.*, vol. 56, no. 3, pp. 741–748, Mar. 2009.

[2] K. Z. Mao, P. Zhao, and P.-H. Tan, “Supervised learning-based cell image segmentation for p53 immunohistochemistry,” *IEEE Trans. Biomed. Eng.*, vol. 53, no. 6, pp. 1153–1163, Jun. 2006.

[3] X. Yang, H. Li, and X. Zhou, “Nuclei segmentation using marker-controlled watershed, tracking using mean-shift, and Kalman filter in time-lapse microscopy,” *IEEE Trans. Circuits Syst. I: Regular Papers*, vol. 53, no. 11, pp. 2405–2414, Nov. 2006.

[4] L. Yang, P. Meer, and D. J. Foran, “Unsupervised segmentation based on robust estimation and color active contour models,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 9, no. 3, pp. 475–486, Sep. 2005.

[5] W. Chen, M. Reiss, and D. J. Foran, “A prototype for unsupervised analysis of tissue microarrays for cancer research and diagnostics,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 8, no. 2, pp. 89–96, Jun. 2004.

[6] S. Schupp, A. Elmoataz, J. Fadili, P. Herlin, and D. Bloyet, “Image segmentation via multiple active contour models and fuzzy clustering with biomedical applications,” in *Proc. IEEE Int. Conf. Natural Comput.*, vol. 1, pp. 622–625, 2000.

[7] X. Zhou, F. Li, J. Yan, and S. T. C. Wong, “A novel cell segmentation method and cell phase identification using markov model,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 2, pp. 152–157, Mar. 2009.

[8] A. Dufour, V. Shinin, S. Tajbakhsh, N. Guillen-Aghion, J.-C. Olivo-Marin, and C. Zimmer, “Segmenting and tracking fluorescent cells in dynamic 3-d microscopy with coupled active surfaces,” *IEEE Trans. Image Process.*, vol. 14, no. 9, pp. 1396–1410, Sep. 2005.

[9] F. Cloppet and A. Boucher, “Segmentation of overlapping/aggregating nuclei cells in biological images,” in *Proc. IEEE Int. Conf. Pattern Recognit.*, 2008, pp. 1–4.

[10] S. Nasr-Isfahani, A. Mirsafian, and A. Masoudi-Nejad, “A new approach for touching cells segmentation,” in *Proc. IEEE Int. Conf. BioMed. Eng. Informat.*, 2008, vol. 1, pp. 816–820.

[11] W. Wang and H. Song, “Cell cluster image segmentation on form analysis,” in *Proc. IEEE Int. Conf. Natural Comput.*, 2007, vol. 4, pp. 833–836.

[12] W. X. Wang, “Binary image segmentation of aggregates based on polygonal approximation and classification of concavities,” *Pattern Recognit.*, vol. 31, no. 10, pp. 1503–1524, 1998.

[13] G. Cong and B. Parvin, “Model-based segmentation of nuclei,” *Pattern Recognit.*, vol. 33, no. 8, pp. 1383–1393, 2000.

[14] S. Kumar, S. H. Ong, S. Ranganath, T. C. Ong, and F. T. Chew, “A rule-based approach for robust clump splitting,” *Pattern Recognit.*, vol. 39, no. 6, pp. 1088–1098, 2006.

[15] H. Zhou and K. Z. Mao, “Adaptive successive erosion-based cell image segmentation for p53 immunohistochemistry in bladder inverted

papilloma," in *Proc. IEEE Ann. Int. Conf. Eng. Med. Biol. Soc.*, 2005, pp. 6484–6487.

[16] T. Jiang, F. Yang, Y. Fan, and D. J. Evans, "A parallel genetic algorithm for cell image segmentation," *Electron. Notes Theor. Comput. Sci.*, vol. 46, pp. 214–224, 2001.

[17] H.-S. Wu, J. Barba, and J. Gil, "A parametric fitting algorithm for segmentation of cell images," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 3, pp. 400–407, Mar. 1998.

[18] P. Soille, *Morphological Image Analysis: Principles and Applications*. Berlin, Germany: Springer-Verlag, 1999.

[19] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-9, no. 1, pp. 62–66, Jan. 1979.

[20] M. Bouguessa, S. Wang, and H. Sun, "An objective approach to cluster validation," *Pattern Recognit. Lett.*, vol. 27, no. 13, pp. 1419–1430, 2006.

[21] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York: Wiley, 2000.

[22] A. Fitzgibbon, M. Pilu, and R. B. Fisher, "Direct least square fitting of ellipses," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 5, pp. 476–480, May 1999.

[23] C. Kim and J.-N. Hwang, "Fast and automatic video object segmentation and tracking for content-based applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 2, pp. 122–129, Feb. 2002.

Chanho Jung received the B.S. and the M.S. degrees in electronic engineering from Sogang University, Seoul, Korea, in 2004 and 2006, respectively. He is currently working toward the Ph.D. degree at the Computational Imaging Laboratory, Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Korea.

From 2006 to 2008, he was a Research Engineer at Digital TV Research Laboratory, LG Electronics, Seoul, Korea. His current research interests include image/video understanding, computer vision, pattern recognition, and image processing.

Changick Kim (SM'09) was born in Seoul, Korea. He received the B.S. degree in electrical engineering from Yonsei University, Seoul, Korea, in 1989, the M.S. degree in electronics and electrical engineering from Pohang University of Science and Technology, Pohang, Korea, in 1991, and the Ph.D. degree in electrical engineering from the University of Washington, Seattle, in 2000, respectively.

From 2000 to 2005, he was a Senior Member of Technical Staff at Epson Research and Development, Inc., Palo Alto, CA. From 2005 to 2009, he was an Associate Professor in the School of Engineering, Information and Communications University, Daejeon, Korea. Since March 2009, he has been with the Department of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon, Korea, where he is currently an Associate Professor. His current research interests include multimedia signal processing, 3-D video processing, image/video understanding, intelligent media processing, and video coding for IPTV.

Seoung Wan Chae was born in Daegu, Korea. He graduated from Hallym University Medical School, in 1991, and received the Ph.D. degree in pathology from Hallym University, Chuncheon, Korea, in 1999.

From 1991 to 1996, he interned and trained in Anatomic Pathology at Chuncheon Sacred Heart Hospital, Chuncheon, Korea. From 1999 to 2000, he was a Fellow of pathology, Seoul National University Hospital, Seoul, Korea. From 2000 to 2002, he was an Instructor of pathology, Hankang Hospital, Hallym University, Seoul, Korea. From 2002 to 2006, he was an Assistant Professor of pathology, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, Seoul, Korea, where he has been an Associate Professor in the Department of Pathology since 2006.His current research interests include gastrointestinal pathology and cellular morphometry.

Sukjoong Oh was born in Seoul, Korea. He graduated from Hanyang University Medical School, Seoul, Korea, in 1992, and received the Ph.D. degree in medical oncology from Ulsan University, Ulsan, Korea, in 2004.

From 1996 to 2000, he trained in Internal Medicine at Hanyang University Hospital, Seoul. From 2000 to 2002, he was a Fellow of Medical Oncology and Hematolgy at Asan Medical Center, Seoul. Since 2003, he has been a member of the medical staff at division of Medical Oncology and Hematology, Department of Internal Medicine, Kangbuk Samsung Hospital and also as an Associate Professor of Internal Medicine, School of Medicine, Sunkyunkwan University, Seoul, Korea. His current research interests are genetics and molecular research in hematologic malignancy.