REPRESENTING CLUMPS OF CELL NUCLEI AS UNIONS OF ELLIPTIC SHAPES BY USING THE MDL PRINCIPLE

Jenni Hukkanen[†], Edmond Sabo[‡], and Ioan Tabus[†]

[†]Department of Signal Processing
Tampere University of Technology
PO Box 553 Tampere, Finland
jenni.hukkanen@tut.fi; ioan.tabus@.tut.fi

[‡]Department of Pathology
Rappaport Faculty of Medicine, Technion
Haifa, Israel
e_sabo@rambam.health.gov.il

ABSTRACT

We discuss the problem of interpreting clumps (or clusters) of nuclei in histological images as unions of elliptical shapes, each ellipse representing one nucleus. The difficult part is to rank various interpretations, involving different numbers of ellipses, and our approach is an information theoretic one where the score for each interpretation is computed using the minimum description length (MDL) principle for a simple parametric family of models. We show how to evaluate MDL for the proposed family using the codelength of an implementable method, which does not involve any asymptotic approximations. We then show how to locally improve the ellipse parameters of a given initial interpretation so that its MDL score is minimized. The initial and final MDL scores of each competing interpretation are then used for deciding which interpretation is the least redundant. We perform a preliminary study involving human subjects for proposing interpretations of the clumps and we also obtain interpretations by an improved version of the existing ellipse fitting algorithm SNEF. We study the variability between the human subject interpretations and compare it with the variability of SNEF algorithm. Finally, the results are examined by a pathology expert for assessing the quality of the MDL based decisions.

1. INTRODUCTION

Histological images are images of thin tissue samples, which are analyzed by expert pathologists for providing medical diagnostic and evaluating the grade of the disease. Every pathologist may analyze daily tens of hematoxylin and eosin (H&E) stained histological images. Each such image may contain hundreds of nuclei and a number of image analysis tasks are implicitly performed by the pathologists for deciding a certain diagnostic: segmentation of nuclei and description of their features like orientation, eccentricity, distance among them. Designing algorithms for solving these image analysis tasks will provide valuable assistance to the pathologist and also constitutes a first step towards automatic diagnosis.

Ideally, if the section of the tissue will be thin enough we will observe in the image only one layer of nuclei, where there will be no overlaps. However, the thickness of the sections is in practice much higher than a single nuclei layer, so that we observe a three-dimensional volume of tissue, whose projection on the bidimensional image will result in overlapping nuclei. Also, diseased tissue will present cells with abnormal size of their nuclei, which are almost touching in the bidimensional H&E image. Hence, a standard segmentation technique applied to the H&E image will face two kinds of tasks: first, segment the well separated nuclei which is easily done even by a simple thresholding operation; second, segment the clump of overlapping nuclei and interpret them into the constituent nuclei. Most segmentation methods will find easily the contours of separated nuclei and the contours of the clumps, but will not be able to provide an interpretation of the clumps in terms of overlapping nuclei.

The interpretation of clumps is difficult because it involves touching and overlapping objects, and identifying nuclei orientations and sizes require fitting overlapping objects, thus is not akin to segmentation. There are a number of techniques proposed for interpreting a clump as a set of overlapping regular shapes. For the separation one may use some prior knowledge about the shape of the objects. In the case of cell nuclei the most typical assumptions about the shapes are convexity (i.e. [1, 2]), ellipticity [3], or both [4]. Algorithms which rely on the assumption that objects are convex generally try to find concavity points from the extracted contours and link them to obtain lines that split clustered objects into individual objects. However the interpretation of overlapping objects is not provided by these methods. Ellipticity assumption will help in case of overlapping and touching objects. Unfortunately, the previous approaches of fitting ellipses relied essentially on the first binary segmentation results, which for most of H&E images are noisy and unreliable. In addition, the information of the gradients inside the clustered objects, which could give important clues of separation lines, is rarely used. The algorithm SNEF, which was recently proposed [5], fits a number of ellipses to the clump, by using jointly gradient and thresholded contour information. SNEF created a large number of candidate ellipses and used a heuristic process for selecting surviving ellipses. It had a number of hard-wired options leading to a unique interpretation of the clump image, which was satisfactory in the clear cases situations analyzed. However, sometimes it is preferable to have at the output of the image analysis task a number of alternative solutions and evaluate by a structure finding method the likelihood of each. After determining the structure with the highest likelihood according to a principled criterion, one may decide on the best possible interpretation, having the least value of criterion, or one may also continue to the next stages of statistical inference with a list of plausible solutions, each being ranked and weighted by its associated criterion value. The goal in this paper is to introduce a principled MDL evaluation scenario for clump interpretations.

We review in Section 2 the algorithm SNEF. In Section 3 we introduce the MDL criterion for ranking several competing interpretations of a clump, based on the codelength of an implementable coding scheme and we present an algorithm for locally improving a given configuration of ellipses so that the MDL score is minimized. In Section 4 we present the experimental setup for evaluating the variability of interpretations given by a number of human subjects and compare it with the variability of the interpretation provided by the SNEF algorithm.

2. A REVIEW OF THE SNEF ALGORITHM

The paper [5] presented an efficient ellipse fitting based algorithm for cell nuclei segmentation from histological H&E stained images. The idea of the algorithm is to estimate the image gradients and to group the connected pixels having high gradient values into a presumptive part of a nuclei contour, to which one ellipse is fitted. The possible discontinuity points between two intersecting ellipses, or alternatively, the groups of contiguous pixels of one ellipse are found by rotating a ray centered at a seed point and picking at each angle a pixel from the high gradient pixels. The obtained pixels are then grouped into connected components, which are grouped in various combinations to be evaluated latter. The process is re-

peated with various grouping of the pixels, resulting in a number of possible ellipses. The algorithm decides based on a heuristic criterion which ellipses to keep in the final interpretation. The initial seeds are determined by a morphological operation and the algorithm tolerates to have more seeds than the real number of nuclei. The algorithm relied on a number of thresholds, like the initial one used for finding the contour of the clump and the one imposed on the gradient values for selecting potential pixels on the ellipses. The algorithm was shown to provide plausible interpretations for clumps of overlapping ellipses. In this paper we allow a number of four distinct choices for the initial thresholds in the algorithm, with two different thresholds for getting the contour and two different thresholds on the gradient image. In this way we get a number of four competing interpretations as the output of the SNEF algorithm, which will be ranked by the MDL based criterion introduced in the next section for principally selecting the best interpretation.

3. RANKING AMONG COMPETING INTERPRETATIONS OF A CLUMP USING THE MDL PRINCIPLE

Minimum description length (MDL) [6] principle provides a principled and systematic framework for comparison between different statistical models (in our case, models for representing geometrical structures). MDL provides a natural trade-off between the complexity of the model and the accuracy of fitting the data.

The MDL principle was previously used for image segmentation in [7], where gray-level images were segmented by minimizing a MDL score of a two-dimensional polynomial model defined on each region, using an optimization technique from the group of continuation methods; the model was refined to include a more precise cost for contours in [8] where MDL costs were used for making decisions during the process of region merging, by which the final segmentation was achieved; finally in [9] a similar technique to [8] is used, operated at a number of different scales and initialized by a mean shift segmentation algorithm.

Differently than in the previously mentioned papers, our problem is not one of segmentation, but one of proposing an interpretation of a region by possibly overlapping ellipses. Therefore our contour costs will have a different form. We present a fully implementable coding algorithm, which provides the codelengths for MDL criterion, as opposed to the papers [7, 8, 9], which use asymptotic expressions of the parameter costs.

The basic idea in [7, 8, 9] and in here is to account for the cost of losslessly describing an image by using the intermediate stage of encoding the segmentation and then encoding the image making use of the already described segmentation. If the segmentation describes accurately the regions of the image, the overall cost will be better than if the segmentation does not describe well the image regions. Using similar notations to [8, 9], we define Ω to be the set of contour pixels, which describe the segmentation, β is the vector of the parameters for the coding distributions. Then we can generically decompose the overall codelength $L(Y,\Omega,\beta)$ in the following terms:

$$L(Y,\Omega,\beta) = L(\Omega) + L(\beta|\Omega) + L(Y|\Omega,\beta), \tag{1}$$

where $L(\Omega)$ is the cost for describing the contour pixels, $L(\beta|\Omega)$ is the cost of describing the coding parameters in each of the regions, and $L(Y|\Omega,\beta)$ is the codelength for encoding the image given the split in regions and using the coding distributions.

3.1 An implementable description of the image

The description of the image is done in a perfectly lossless way by specifying the following: the parameters of the ellipses, the procedure for constructing the contour of the nuclei cluster given the ellipses, the parametric description of the interior and of the exterior of the nuclei cluster and finally the residuals for all pixels in the image. Our hypothesis is that each ellipse represents a nucleus, which may partially occlude (or overlap with) other nuclei.

Codelength $L(\Omega)$ for representing the ellipses

We define Ω as the contour pixels, forming the outer boundary of the union of the n_E ellipses used in a given interpretation. This set can be obtained by describing the n_E ellipses and then forming their union and taking the boundary set of it. Each ellipse is represented and encoded by using the parametrization having the parameter vector $\alpha = [x_0\ y_0\ a\ b\ \theta]$, where (x_0, y_0) are the coordinates of the center of ellipse, a is the major axis, b is the minor axis, and θ is the angle between the x axis and the major axis of the ellipse. For an image with n_r rows and n_c columns, the possible range of the parameters is from 0 to the following maximum values: $\alpha_{Max,1} = n_r$, $\alpha_{Max,2} = n_c$, $\alpha_{Max,3} = \alpha_{Max,4} = \sqrt{n_c^2 + n_r^2}$, $\alpha_{Max,5} = \pi$. Encoding of the parameter α_i is realized by uniformly quantizing its value by using 2^b reconstruction levels in the range $(0, \alpha_{Max,i})$, with a cost of b bits per parameter. The resolution in the parameter space is $\Delta_{\alpha_i} = \alpha_{Max,i}/2^b$. Experimentally we found that $b = 7$ is providing the best overall codelength. Thus each ellipse requires $5b = 35$ bits for encoding its parameters and we need a total codelength $L(\Omega) = 5bn_E$ for all n_E ellipses needed to represent Ω.

Codelength $L(\beta|\Omega)$ for representing the parameters of coding distributions

For the description of the foreground and background of the image one can use polynomial functions as in [7, 8, 9]. However, we observed that a constant model for the foreground (with a constant μ_F) and a constant model for the background (with a constant μ_B) are providing good segmentations. Since the significance of the constant level is that of a gray level in the luminance image, we require 8 bits for each of the two constant levels μ_F and μ_B. The residual image is encoded using Golomb-Rice codes for a doubly exponential distribution, which is known to be very efficient way for lossless encoding of images [10]. We will use two different coding distributions, one for the foreground and one for the background, requiring to specify two Golomb-Rice coding parameters, ℓ_F and ℓ_B, each of 8 bits, thus in total $L(\beta|\Omega) = 32$ bits.

Codelength $L(Y|\Omega,\beta)$ for representing the image given the ellipses and coding parameters

Our algorithm operates in the luminance domain of the original H&E image (as most steps of SNEF algorithm operate also on the luminance component). We arrange columnwise the luminance values $Y(i,j)$ falling inside the contour Ω in a long vector $\{z_k, k = 1,\ldots,n_F\}$ and define the residuals as $\varepsilon_i = z_i - \hat{z}_i$, where $\hat{z}_i = \mu_F$. Assuming a two sided exponential distribution $P(\varepsilon) = C_\lambda e^{-|\varepsilon|/\lambda}$ for the residuals ε_i, the best codes are obtained by applying first a mapping of signed errors to unsigned errors by

$$\gamma_i = \begin{cases} 2\varepsilon_i & if \quad \varepsilon_i \geq 0 \\ 2|\varepsilon_i| - 1 & if \quad \varepsilon_i < 0 \end{cases} . \tag{2}$$

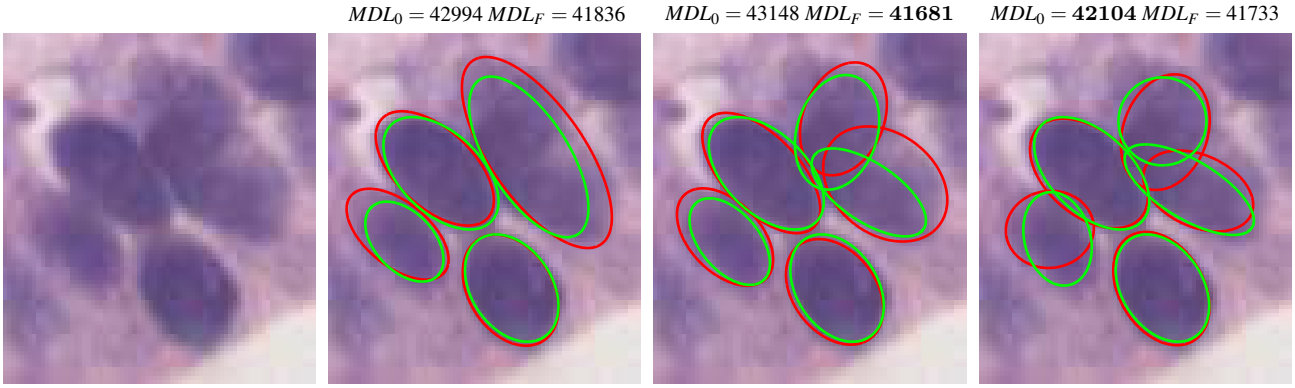
The Golomb-Rice code of parameter $k_F = 2^{\ell_F}$ will encode unary the value $\lfloor \frac{\gamma_i}{2^{\ell_F}} \rfloor$ and then will transmit the reminder $\gamma_i - 2^{\ell_F} \lfloor \frac{\gamma_i}{2^{\ell_F}} \rfloor$ using ℓ_F bits, requiring in total $\lfloor \frac{\gamma_i}{2^{\ell_F}} \rfloor + \ell_F + 1$ bits. Thus to encode the residuals of the foreground we need

$$L_F \quad = \quad n_F(\ell_F + 1) + \sum_{i=1}^{n_F} \left\lfloor \frac{\gamma_i}{2^{\ell_F}} \right\rfloor . \tag{3}$$

In an identical manner we can derive the codelength L_B for representing the residuals from the background. Thus the overall codelength for both foreground and background pixels is $L(Y|\Omega,\beta) = L_F + L_B$.

3.2 Optimization of MDL criterion

The discrete nature of the parameter space used in our implementable coding scheme suggests a direct optimization of the MDL cost by a relaxation method, where at each iteration one ellipse is selected to be changed, while all others are kept fixed. Changing the parameters of the selected ellipse $\mathscr{E}_i$ is done by letting its vector of parameters to run through the set of vectors obtained by the Cartesian product $\bigotimes_{i=1}^{5} \{(\alpha_i - \lambda\Delta_{\alpha_i}), \alpha_i, (\alpha_i + \lambda\Delta_{\alpha_i})\}$ and thus at one iteration of the relaxation process we evaluate $L(Y,\Omega,\beta)$ a number



(a) (b) $K = 4$ (c) $K = 5$ (d) $K = 5$

Figure 1: Initial ellipses (in red) and their corresponding criteria, MDL_0, and final ellipses (in green), after the iterative local optimization and their corresponding criteria, MDL_F. (a) Original RGB image; (b) Interpretation C_1 of one subject; (c) Interpretation C_2 of the same subject; (d) SNEF interpretation. Lowest MDL is obtained for the interpretation with $K = 5$ ellipses given by the human subject, closely followed by the interpretation of SNEF, also for $K = 5$ ellipses.

of 3^5 times and keep in the end the parameters of $\mathscr{E}_i$ that provided the lowest codelength. The evaluation process can be organized in an efficient manner, since a lot of computations can be easily updated for the next evaluation.

One relaxation cycle ends after we went and changed all the n_E ellipses. In the experiments reported here we had a number of 5 relaxation cycles, where the parameters λ were taken in turn from the list $[1,2,1,2,1]$.

4. EXPERIMENTAL RESULTS

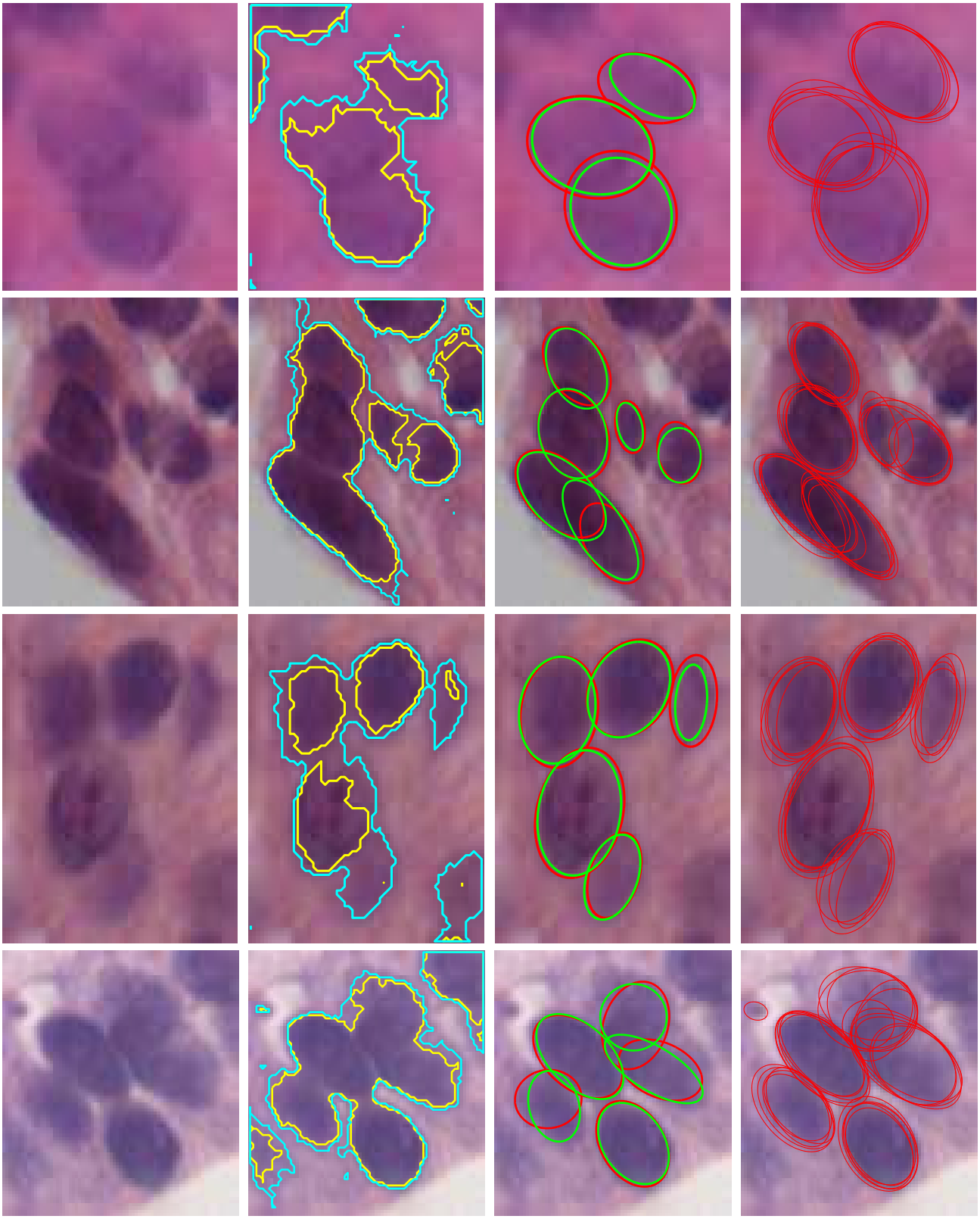
We selected from a set of histological images a number of $n_I = 24$ clumps. We collected interpretations from $n_S = 5$ subjects, who were instructed to do the following: use an interactive graphical routine to mark on the image the contour of all shapes that are resembling (even slightly) an elongated shape not touching the border of the image (our clumps were all selected for simplicity to be fully included in the analysis window). The best fitting ellipse to each contour was then computed by a constrained least square fitting and presented to the subject overlapped on the original image; the interactive graphical routine also allowed the subject to adjust the shape of each ellipse to reach the best fit, judged subjectively. The ellipses were labeled on the screen and the subjects continued by delivering a number of likely explanations of the image I_i, e.g. a subject S_k can provide $n_C(k) = 2$ configurations: $C_1 = \{\mathscr{E}_1, \mathscr{E}_2, \mathscr{E}_3\}$ and $C_2 = \{\mathscr{E}_1, \mathscr{E}_2\}$. A degree of belief $\hat{p}(I_i, S_k, C_\ell)$ in the configuration explanations was also input from each subject S_k.

In Figure 1 we show the type of inference, which is allowed by using the MDL structure selection: in Figures 1 (b)-(d) we have three interpretations of the clump from Figure 1 a), two provided by a human subject and one provided by the SNEF algorithm. The initial ellipses are shown in red and the corresponding MDL values are 42994, 43148 and 42104, with the lowest initial MDL provided by SNEF algorithm, for an interpretation with 5 ellipses. It is interesting that the MDL_0 values for (b) and (c) favor $K = 4$ ellipses. However, after running our locally optimization algorithm we get a consistent result, both interpretations with $K = 5$ achieve the lowest value of MDL_F. The percentage of MDL reduction is $100(MDL_0 - MDL_F)/MDL_0$ and the corresponding values are 2.69%, 3.40%, and 0.88%.

Next we illustrate in detail the type of variability of resulting segmentations, when starting from different human subjective evaluations of ellipse interpretations, and when using the ellipse fitting algorithm. We have run the algorithm SNEF four times, with two different thresholds for getting the contour and two different thresholds on the gradient image. For finding the threshold values for the luminance image we used dual thresholding [11] and Otsu thresholding [12]. Otsu thresholding minimizes the inter class variance, while dual thresholding is developed for histological images consisting of nuclei, cytoplasm, and background region. For the gradient image we use two thresholds: once Otsu threshold and second time we completely ignored the gradient image. This led to final segmentations involving different number of ellipses. We show in Figure 2 the results obtained for the 4 clump images from column (a) of Figure 2. The optimal contours of the clumps obtained by optimal Otsu threshold and by dual thresholding are presented in the column (b) of Figure 2. In column (c) of Figure 2 we show the best results (in terms of MDL) of the SNEF algorithm, for the initial configuration and also after locally adjusting the parameters of the ellipses by using the proposed iterative algorithm. Finally, in column (d) of Figure 2 we show all ellipses traced by the human subjects. We note that the clumps in rows 2 and 4 presented are situations more difficult to interpret. The difficulty of the task comes from a number of competing interpretations of some areas of the clumps: elongated shapes that can be split in two or not; light contrast shapes, which one may even consider to be only an artifact due to noise; very poor contours, due to fading of the intensity. Overall SNEF provided plausible interpretations, comparable to those of the human subjects.

This process has created a number of multiple interpretations for each of the 24 images. We define for the time being the "ground truth" of a given feature in one given image as the average of that given quantity over all interpretations provided by subjects (not including the SNEF algorithm) for that image. For example we can introduce the average $\overline{n}_E(I_i)$ of the number of ellipses found for image I_i as $\overline{n}_E(I_i) = \frac{1}{n_S}\sum_{k=1}^{n_S}\sum_{\ell=1}^{n_C(k)} \hat{p}(I_i, S_k, C_\ell) n_E(I_i, S_k, C_\ell)$ and similarly define the variance of the number of ellipses n_E as $\sigma^2(n_E(I_i)) = \frac{1}{n_S}\sum_k \sum_\ell \hat{p}(I_i, S_k, C_\ell)(n_E(I_i, S_k, C_\ell) - \overline{n}_E(I_i))^2$. We define in this way averages and variances over each image for the following quantities: initial MDL_0 value of subject's configurations, final MDL value, MDL_F, after adjusting each configuration towards a lower MDL. Then we evaluate similarly deviations of the results obtained by SNEF algorithm against "the ground truth", e.g., $\Delta(n_E(I_i)) = |n_E(I_i, SNEF) - \overline{n}_E(I_i)|$. We present in Figure 3 (a) the standard deviations $\sigma(n_E(I_i))$, for each image I_i and the values $\Delta(n_E(I_i))$. Similarly the Figures 3 (b) and 3 (c) show the variability observed in the initial values MDL_0 and in the final values MDL_F.

In Table 1 we present also another view at the variability in the decisions regarding the number of ellipses. For each image we can express the degree of belief of all n_S subjects on a given number n of ellipses by $P(n, I_i) = \frac{1}{n_S}\sum_k \sum_{\ell | n_E(I_i, S_k, C_\ell) = n} P(I_i, S_k, C_\ell)$. We redefine for each image the ground truth as the maximum likelihood



(a) (b) (c) (d)

Figure 2: (a) Original RGB images. (b) Boundaries obtained by Otsu thresholding (cyan) and dual thresholding (yellow) superposed over the original image. (c) Best SNEF results: initial configuration (red) and after iterative improvement of MDL (green). (d) All ellipses traced by the 5 human subjects.

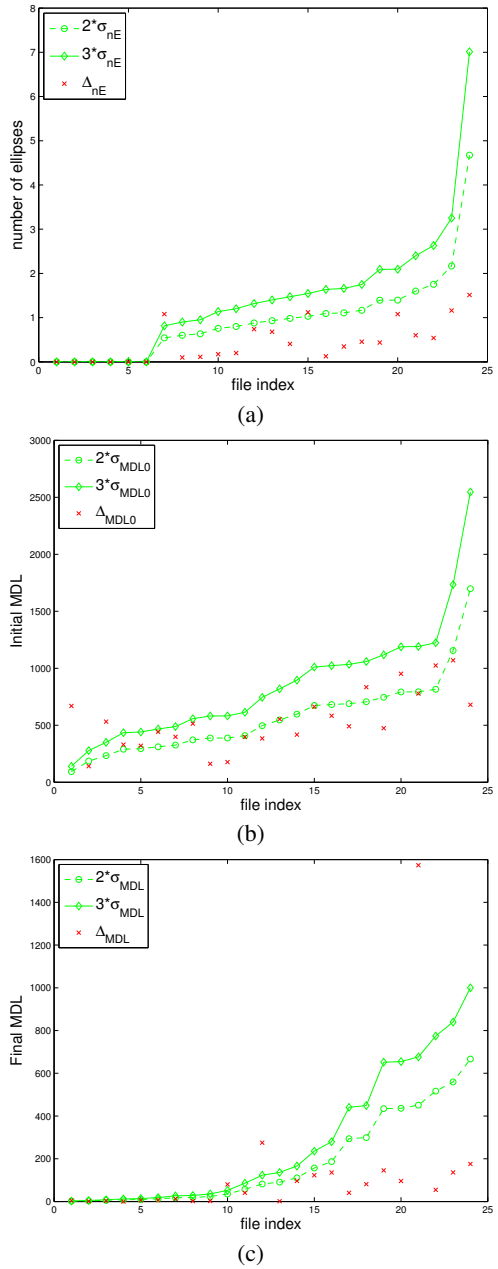


Figure 3: Variability of estimated number of ellipses and of MDL before and after the iterative algorithm.

Δn_E	-2	-1	0	1	2
$\hat{P}(\Delta n_E)$	0.0108	0.0625	0.7670	0.1206	0.0307
$\hat{P}_{SNEF}(\Delta n_E)$	0	0.1667	0.7083	0.1250	0

Table 1: Probabilities of making Δn_E mistakes in the number of ellipses, by the human subjects ($\hat{P}(\Delta n_E)$) and by SNEF algorithm ($\hat{P}_{SNEF}(\Delta n_E)$).

of the values n_E proposed by each subject, by finding the value $\hat{n}_E(I_i)$, which maximizes $P(n, I_i)$. We define the probabilities of the errors over all images as $\hat{P}(\Delta n_E) = \frac{1}{n_I}\sum_i P(\hat{n}_E(I_i) + \Delta n_E, I_i)$. We also define a similar quantity for the $n_{E,SNEF}(I_i)$ values provided by the SNEF algorithm, $\hat{P}_{SNEF}(\Delta n_E) = \frac{1}{n_I}\sum_i P(n_{E,SNEF}(I_i) = \Delta n_E + \hat{n}_E(I_i))$.

5. DISCUSSIONS AND CONCLUSIONS

Before seeing the MDL results and the segmentations of Figure 1, the pathology expert decided that the interpretation with $K = 5$ is the best interpretation, in conclusion MDL evaluation for Figure 1 provides very good results, in agreement with the pathology expert.

From Table 1 we can note that the probability of finding the right number of ellipses is a little higher for human subjects than for SNEF algorithm, while the probabilities of mistakes spread over a higher number of mistake numbers for human than for SNEF algorithm.

From Figure 3 we see that the variability of the MDL criterion over the provided human interpretations is much higher than that of the final values MDL_F, obtained with our iterative algorithm. Also we note that the deviations Δ_{MDL_F} from the ground truth of the SNEF algorithm are in general lower than two times the standard deviation of the human subjects obtained MDL.

REFERENCES

[1] S. Kumar, S.H. Ong, S. Ranganath, T.C. Ong, and F.T. Chew, "A rule-based approach for robust clump splitting," *Pattern Recognition*, vol. 39, pp. 1088–1098, 2006.

[2] M. Faessel and F. Courtois, "Touching grain kernels separation by gap-filling," *Image Anal Stereol*, vol. 28, pp. 195–203, 2009.

[3] G. Zhang, D.S. Jayas, and N.D.G. White, "Separation of touching grain kernels in an image by ellipse fitting algorithm," *Biosystems Engineering*, vol. 92, pp. 135–142, 2005.

[4] X. Bai, C. Sun, and F. Zhou, "Splitting touching cells based on concave points and ellipse fitting," *Pattern Recognition*, vol. 42, pp. 2434–2446, 2009.

[5] J. Hukkanen, A. Hategan, E. Sabo, and I. Tabus, "Segmentation of cell nuclei from histological images by ellipse fitting," in *Proc. of European Signal Processing Conference*, Aalborg, Denmark, August 2010, pp. 1219–1223.

[6] J. Rissanen, "Modeling by shortest data description," *Automatica*, vol. 14, pp. 465–471, 1978.

[7] Y.G. Leclerc, "Constructing simple stable descriptions for image partitioning," *International Jounal of Computer Vision*, vol. 3, pp. 73–102, 1989.

[8] T. Kanungo, B. Dom, W. Niblack, and D. Steele, "A fast algorithm for MDL-based multi-band image segmentation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, June 1994, pp. 609–616.

[9] Q. Luo and T.M. Khoshgoftaar, "Unsupervised multiscale color image segmentation based on MDL principle," *IEEE Trans. on Image Processing*, vol. 15, pp. 2755–2761, 2006.

[10] M.J. Weinberger, G. Seroussi, and G. Sapiro, "LOCO-I: A low complexity, context-based, lossless image compression algorithm," in *Proc. of the IEEE Data Compression Conference*, Snowbird, Utah, March 1996.

[11] M. Hu, X. Ping, and Y. Ding, "Automated cell nucleus segmentation using improved snake," in *Proc. of International Conference on Image Processing*, Singapore, October 2004, pp. 2737–2740.

[12] N. Otsu, "A threshold selection method from gray level histograms," *IEEE Trans. Systems, Man ans Cybernetics*, vol. 9, pp. 62–66, 1979.