Automatic Clump Splitting for Cell Quantification in Microscopical Images

Gloria Díaz, Fabio Gonzalez, and Eduardo Romero

Bioingenium Research Group, National University of Colombia,
Bogotá, Colombia
{gmdiazc,fagonzalezo,edromero}@unal.edu.co
http://www.bioingenium.unal.edu.co

Abstract. This paper presents an original method for splitting overlapped cells in microscopical images, based on a template matching strategy. First, a single template cell is estimated using an Expectation Maximization algorithm applied to a collection of correctly segmented cells from the original image. Next, a process based on matching the template against the clumped shape and removing the matched area is applied iteratively. A chain code representation is used for establishing best correlation between these two shapes. Maximal correlation point is used as an landmark point for the registration approach, which finds the affine transformation that maximises the intersection area between both shapes. Evaluation was carried out on 18 images in which 52 clumped shapes were present. The number of found cells was compared with the number of cells counted by an expert and results show agreement on a 93 % of the cases.

Keywords: Cell quantification, Overlapping objects, Segmentation, Clump splitting.

1 Introduction

Clumping of objects of interest is a relatively frequent phenomenon in different computer vision domains. Its identification results crucial in many cytological applications [1,2,3], in which the expected result is a population count; although human experts are capable of separating the constituent objects, most real applications require a count of a large number of these objects, thereby many conclusions of cytological studies lye on statistical or qualitative approaches [4]. Manual methods have been replaced in hematological cell counting by automated techniques because of a superior repeatability and the avoidance of the many error sources present in manual methods [4]. Besides, manual strategies are in general limited in cases such as random aggregates of cells produced by smearing failures or dye deterioration [4].

Available clump splitting methods are based on prior knowledge about shape, size or region gray level intensities [5,6,7]. These methods include mathematical morphology [3,8,9], watershed techniques [10,11] and concavity analysis [12,13,3].

L. Rueda, D. Mery, and J. Kittler (Eds.): CIARP 2007, LNCS 4756, pp. 763–772, 2007.

Di Ruberto [8] applies a size defined disk as a structural element to separate clumped red cells while Ross [9] complements it using a gray level granulometry for separating objects in the image. Concavity analysis methods are based on the hypothesis that superimposed objects can be separated at some specific cut points in which either the curvature abruptly changes or the overlapped objects present differences in the gray level intensities. The drawback of these methods is that they are only applicable for objects with specific shapes and sizes. On the other hand, Kumar [3] proposes a method based on a concavity analysis, adaptable to many shapes and sizes and which depends on a set of parameters that are obtained from a large set of training samples. However, this method is not accurate enough (79 %), many samples are synthetic and there is not a study of the degree of overlapping at which the method is capable to deal with.

The clump-splitting method herein proposed addresses the issue that for the particular case of cytology, the *a priori* information about the predominant cell shape and size are already present in the image. For this, a cell model obtained from the image is used for separating cell aggregates. This approach is simple and permits reliable quantification, independent of any pre-determined geometric feature (shape and size). It enables the accurate splitting of clumps composed of cells of different sizes and with a variable degree of overlap. This paper is organized as follows: the construction of a cell model template, estimated from single cells segmented from actual microscopical images is presented in Section 2.2. This template is then used for an efficient search of similar objects in the clumped shapes via a template matching approach, (Section 2.3). Finally, some preliminary results and conclusions are presented in Sections 3 and 4 respectively.

2 Methodology

In figure 1 the main steps of the whole process are illustrated. Firstly, single and clumped cells are extracted from an initial image. Then, single cells are used for estimating a cell model, an estimation which is formulated as a maximum likelihood problem and solved with an Expectation Maximization algorithm. The cell model is finally used as a template for splitting cells in clumped shapes. This approach searches the better matching between a chain code representation of the contours of the clumped shape and the cell model.

2.1 Single Cell Extraction

Cell features are highlighted using very specific histological procedures, which mostly consist in coloring the different cell components so that color is essentially the base of any differential diagnosis and the main strategy for finding objects in histological samples [14]. Single cell extraction can be achieved through a variety of segmentation techniques [15]. Cells are herein extracted from a binary partition of the image, obtained from a process in which objects are segmented

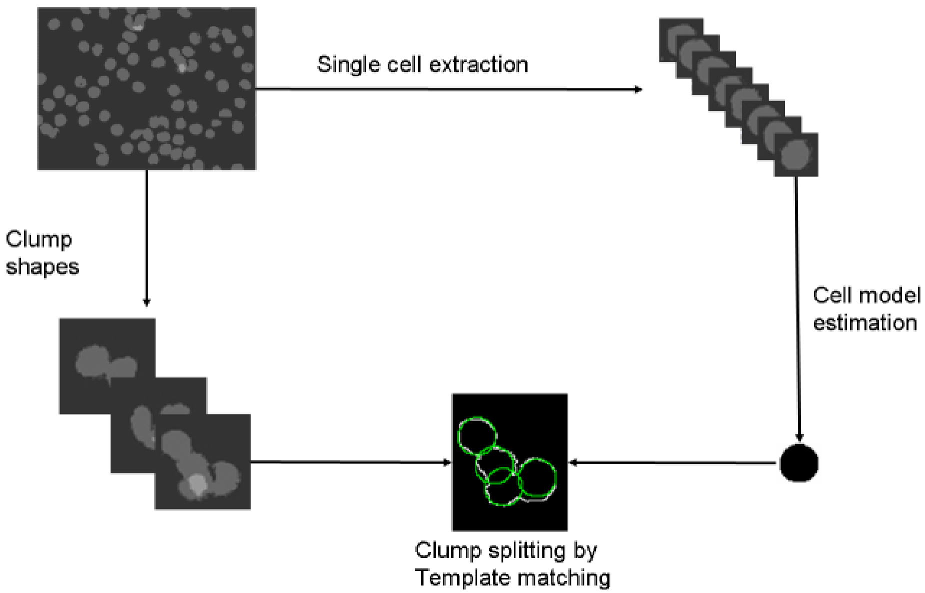

Fig. 1. Proposed method: single and clumped cells are extracted from a initial segmentation. Single cells are then used for estimating a cell model, which is used for splitting clumped shapes via a template matching strategy.

using a color strategy. Therefore, searched objects in these histological images are clustered using their color characteristics.

The problem of color segmentation can be formulated as to find the set of boundaries in the *RGB* cube, which optimally separates tissues. This corresponds to assign to each image pixel a particular class, based on the color structure of the image. Colour classification at the level of pixel is thus the first step for identifying fundamental relationships in the digital image. Evaluation images were segmented using a trained neural network, a multilayer perceptron with one hidden layer, which classified pixels using the RGB cube as the parameter space. Training points were selected by a pathologist from one image and applied to the whole set of histological images. It was needed two training sets, one drawn from images of malaria and the other from plasmocytoma images.

Once pixels are separated into their constituent classes, they are assembled together into objects using neighbor information. Formally, this is a connected operator graph [16], which uses filtering operations for finding relevant morphological structures. This image representation easily permits separation of the single and clumped objects in the image. The graph is constructed with the number of levels needed to represent the hierarchical relationships of the image. Once the graph representation is complete, a number of connected operators are then successively applied for removing redundant information and identifying interest objects. Finally, single cells are extracted and aligned into the same axis using a standard principal component analysis (PCA) [15] and the single cells bounding boxes dimensions are set to the bounding box of the larger feature.

2.2 Template Construction Via the EM Algorithm

For the cell template construction, we assume that each single cell drawn from the image is one instance of a true model. Each is assumed to be generated from a process that modifies the true model by adding a random noise, which models the complex interaction of factors such as the biological variability, the histological procedure and the illumination capturing conditions.

Let $D_i = (D_i^1, \ldots, D_i^n)$ be a vector of n elements, which stores the n binary pixel values of a single cell image, with $i = 1 \ldots N$ and N the number of single cells extracted from the image. Let I be a vector of n elements too, which stands for the pixel values of the ideal cell (true model) so that

$$D_i^j = T_i(I^j) \tag{1}$$

where T is a stochastic function that generates the model instance and is defined as follows

$$T_i(1) = \begin{cases} 1 \text{ with probability } p_i \\ 0 \text{ with probability } 1 - p_i \end{cases} \qquad T_i(0) = \begin{cases} 1 \text{ with probability } 1 - q_i \\ 0 \text{ with probability } q_i \end{cases} \tag{2}$$

Where p_i and q_i control the probability of error on the generated instance. A T_i with $p_i = q_i = 1$ means that instances generated by T_i corresponds to the true model. The problem is then to find the p_i and q_i values which maximise the likelihood of the instances being generated from the model:

$$(p, q, I) = \arg\max_{p,q,I}(L(D|p, q, I)) \tag{3}$$

where the likelihood

$$\begin{aligned} L(D|p, q, I) &= \prod_{i=1}^{N} \prod_{j=1}^{n} P(D_i^j | p_i, q_i, I^j) & (4) \\ &= \prod_{i=1}^{N} \prod_{j=1}^{n} P(D_i^j | p_i, q_i, I^j = 1)^{I^j} P(D_i^j | p_i, q_i, I^j = 0)^{1-I^j} & (5) \\ &= \prod_{i=1}^{N} \prod_{j=1}^{n} p_i^{I^j D_i^j} q_i^{(1-I^j)(1-D_i^j)} (1 - p_i)^{I^j (1-D_i^j)} (1 - q_i)^{(1-I^j) D_i^j} & (6) \end{aligned}$$

A first naive approximation to this problem could be an intensive search of the parameters, but this is no feasible because of the size of the parameter space, which is potentially infinite. An alternative approach is to iteratively improve the estimation of the optimal parameters. For this purpose, a Expectation Maximization (EM) strategy was adapted from the original work of Warfield [17].

The main idea of the approach is to consider the true model (I) as a hidden variable, which is estimated from the observed data and a set of values for the parameters p_i and q_i. The initial values of p_i and q_i are further improved by local optimization. The process of alternatively estimate I (expectation step) and improve the p_i and q_i values (maximization step) is iterated until convergence. This convergence is guaranteed since the likelihood function has an upper bound, as was stated in [18].

The initial parameter estimates p_i and q_i are set to 0.9, as the fundamental hypothesis in this work is that the instances do not differ too much from the true model. The final estimation of I corresponds to the true model that will be later used as a template to find cells in the input image.

2.3 Splitting Via Template Matching Strategy

Tradionally, template matching techniques have been considered as expensive regarding computational resources since the template must slide over whole image. However, the approach herein used is mainly based on a simplified version of both the template and the clumped shape through a chain code representation, which searches for an anchorage point that results in a "best match" when the two shapes are superimposed.

A chain code is typically used to represent the object boundary by a sequence of straight-line segments with their associated directions. A randomly selected pixel from the object boundary is chosen as the initial point. Afterwards, the pixel's neighbors are numbered from 0 to 7 (8-neighbor mask) and the pixels belonging to the boundary are selected following a clockwise direction. Finally, the obtained chain code is normalized for achieving an invariant representations regarding the initial point and orientation [15]. This normalization is performed computing the distance difference between two consecutive segments and assuming that the chain code is a circular sequence.

Once a chain code representation is achieved for both the clumped and template shapes, a maximal correlation point is determined in the registration phase. This point is from now on a landmark which limits transformations of the found template shape. Provided that our true model may differ from cells which result trapped into aggregates and which generally are deformed because of the contact with other cells, this landmark is used to bond both the ideal model contour and the clump boundary and constitutes the initial search point. Registration is addressed to find the affine transformation which maximises the intersected areas between the two shapes: the template and the clumped. Overall, the template size (width and height) was varied from 70% to 120% for allowing to find a "best match", even if the cell was deformed into the clump. Likewise, orientation was varied in steps of 5%, sliding the template code over the clumped shape. After a first cell is found, its corresponding intersection surface is eliminated of the clumped shape as well as its equivalency from the chain code. Procedure is iterated until the remaining area is lower than 0.2 of the original clumpled shape.

3 Experimentation

3.1 Experimental Setup

In the present investigation we performed evaluations on two different types of cells. Figure 2 displays two microscopical images obtained from the two cell types: plasmocytoma (left panel) and thin blood smears infected with malaria parasite (right panel). Upper row displays the original digital images, while bottom row shows the obtained images using the segmentation approach described before. Our objective was thus to find the cells within the clumped shapes, formed after the segmentation process.

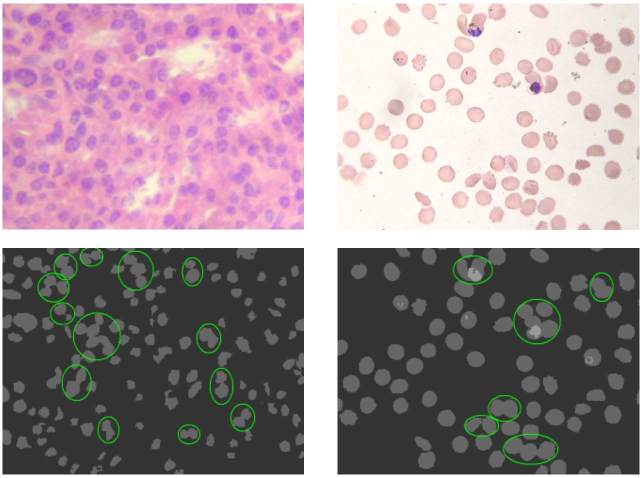

Fig. 2. Fist row corresponds to the original microscopical images. Bottom row displays segmented images obtained from the original ones. Several clumped shapes appear in both cases, a result of the overlapped cells.

A group of 18 microscopical images was used for evaluation, 14 from thin blood stained samples and 4 from a plasmocytoma slide, chosen from two different unrelated studies. These samples corresponded to a two very different tissues, each entailed with different color properties.

3.2 Results and Discussion

Figure 3 shows the final and intermediate results. Upper row (First row) displays, from left to right different microscopical images, among which the first two are extracted from thin healthy blood samples and the next five are extracted from thin blood samples infected with Plasmodium falciparum. In the same row the last five images come from a plasmocytoma slide, characterized by large nuclei with different shapes, sizes and in which the variable to determine is the number of nuclei.

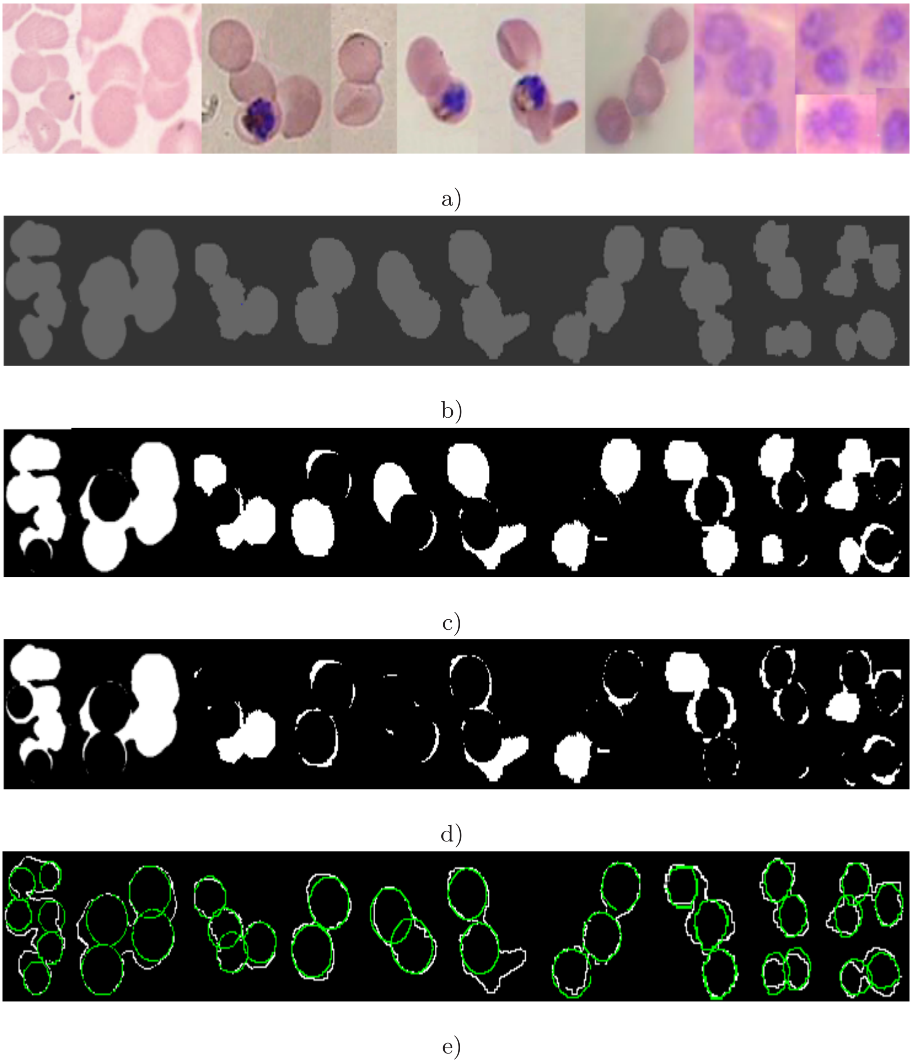

Fig. 3. Figure illustrates the whole process using actual cytological images from different tissues. From the upper to the lower row: row a) displays the original digital images and the first five images from left to right correspond to red cells infected by *plasmodium falciparum*; the rest of the row shows images from a plasmocytoma, a kind of cancer in the lymphatic system. Row b) depicts the binarized images after the color classifier has segmented objects, rows c) and d) show first and second iterations of the proposed method and finally, row e) shows the superimposed results of the splitting cells and the clumped contour shape.

From upper to lower row, results shown in Figure 3 summarise the entire process: row *a* illustrates some examples of sets of cells which are touching or overlapping each other in the two cases herein evaluated (thin blood stains and plasmocytoma). Row *b* shows results after the binarization strategy for every original image in the upper row. It should be strengthen out that at this state, the graph has been already constructed and every single cell has been ruled out so that the graph is uniquely composed of clumped shapes. Notice that the color strategy can also produce overlapping shapes because of the segmentation process, see for example the eighth panel (from left to right) of row *a* and observe the resulted segmentation at the corresponding image in row *b*. Overall, cells are easily separated using color differences. However, the segmentation process may result in complex shapes such as the shown in the mentioned panel. For this reason, yet color characteristics are at the base of differences among objects, they are difficult to establish since histological objects are complex mixes of different intensities and chrominances which are seen in the RGB space as boundaries varying from one image to other. Rows *c* and *d* illustrate the splitting process i.e. a first best matching is shown in row *c* while a second best matching is displayed in row *d*. Observe that there is no a systematic trend about a preferred initial location among the whole set of assessed shapes. Finally, row *e* shows the original clumped contour superimposed with the different locations at which the template has found a relevant shape.

The proposed technique was applied to the set of evaluation images, the identified cells were quantified and the results compared against a manual quantification. In every case, the algorithm was able to match a shape which definitely was an actual cell, a finding which was correlated with the results obtained from observations performed by an expert on the whole set of images. Automatic quantification (the number of found cells for these shapes) coincided in 49 of 52 clumped shapes, resulting in a 93% agreement. Failures were mostly due to an overlapping larger than 50% or to very deformed cells which have lost their geometrical properties and were very different from the estimated template. Regarding time performance, the whole process for a 640 × 480 image size was $0.7 \pm 0.17\,s$.

4 Conclusions

Automatic methods for performing a precise cell counting are limited by a large number of artifacts, among which the formation of clumped shapes is one of the most frequent. In this research, an entirely automatic method is proposed for splitting cells within clumped shapes. The process starts by performing a binarization of the microscopical image, after which every single cell is counted and stored for the construction of a model cell. This cell model is inferred from single cells by an Expectation Maximization algorithm applied at the level of each pixel. The clumped and template contours are then transformed into a chain code, which is used for the registration phase. Registering is performed through affine transformations of the template, under the restriction that the maximal correlation point between the two shapes is fixed. The proposed method has shown to be robust by splitting cells of diverse sizes and shapes whose overlap varies, it is also reliable and reproducible on the test group of evaluation images.

Future work includes the evaluation of the proposed method in different applications domains and the exploration of different representation alternatives for the true cell model.

Acknowledgments. This work is partially supported by a grant from the Colombian Institute for the Advancement of Science and Technology (COLCIENCIAS), Grant no. 109-2005. Smears used in this study was supplied by Research Groups in Bioactive Principles in Medicinal Plants and Molecular Pathology of National University of Colombia.

References

1. Ong, S.H., Jayasooriah, Y.H.H., Sinniah, R.: Decomposition of digital clumps into convex parts by contour tracing and labelling. Pattern Recognition Letters 13, 789–795 (1992)
2. Bowie, J., Young, I.: An analysis technique for biological shape-ii. Acta Cytologica 21, 455–464 (1977)
3. Kumar, S., Ong, S., Ranganath, S., Ong, T., Chew, F.: Arule based approach for robust clump splitting. Pattern Recognition 39, 1088–1098 (2006)
4. Hall, R., Malia, R.G.: Medical Laboratory Haematology, 2nd edn. Butterworth-Heinemann Ltd. (1991)
5. Wang, Z., Ben-Arie, J.: Model based segmentation and detection of affine transformed shapes in cluttered images. In: International Conference on Image Processing (1998)
6. Gadkari, M.S., Refai, H.H., Sluss, J.J., Broughan, T.A., Broughan, T.A., Naukam, R.: The detection of single hepatocytes within clusters in microscopic images. In: Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation (2004)
7. Wurflinger, T., Stockhausen, J., Meyer-Ebrecht, D., Bocking, A.: Robust automatic coregistration, segmentation, and classification of cell nuclei in multimodal cytopathological microscopic images. Computerized Medical Imaging and Graphics 28, 87–98 (2004)
8. di Ruberto, C., Dempster, A., Khan, S., Jarra, B.: Analysis of infected blood cell images using morphological operators. Image and Vision Computing 20(2), 133–146 (2002)
9. Ross, N.E., Pritchard, C.J., Rubin, D.M., Dusï, A.G.: Automated image processing method for the diagnosis and classification of malaria on thin blood smears. Medical and Biological Engineering and Computing 44, 427–436 (2006)
10. Nilsson, B., Heyden, A.: Segmentation of dense leukocyte clusters. In: Proceedings of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (2001)
11. Althoff, K.: Implementation and evaluation of stem cell segmentation techniques. Technical report, Chalmers University of Technology (2003)
12. Liu, L., Sclaroff, S.: Deformable model guided region split and merge of image regions. Image and Vision Computing 22, 343–354 (2004)
13. Poon, S.S.S., Ward, R.K., Palcic, B.: Automated image detection and segmentation in blood smears. Cytometry 13, 766–774 (1992)

14. Junqueira, L.C., Carneiro, J.: Basic Histology, 10th edn. MacGraw Hill, New York (2003)
15. Gonzlez, R.C., Woods, R.E.: Digital Image Processing. Prentice-Hall, Englewood Cliffs (2002)
16. Braga-Neto, U.: Multiscale connected operators. Journal of Mathematical Imaging and Vision 22, 199–216 (2005)
17. Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (staple): An algorithm for the validation of image segmentation. IEEE Transactions On Medical Imaging 23, 903–921 (2004)
18. Dempster, A., Laird, N., Rubin, D.: Maximum-likelihood form incomplete data via the em algorithm. Journal of Royal Statistical Society 39, 34–37 (1977)