Towards Safer, Faster Prenatal Genetic Tests: Novel Unsupervised, Automatic and Robust Methods of Segmentation of Nuclei and Probes

Christophe Restif

Department of Computing, Oxford Brookes University, Oxford OX33 1HX, UK
christophe.restif@centraliens.net

Abstract. In this paper we present two new methods of segmentation that we developed for nuclei and chromosomic probes – core objects for cytometry medical imaging. Our nucleic segmentation method is mathematically grounded on a novel parametric model of an image histogram, which accounts at the same time for the background noise, the nucleic textures and the nuclei's alterations to the background. We adapted an Expectation-Maximisation algorithm to adjust this model to the histograms of each image and subregion, in a coarse-to-fine approach. The probe segmentation uses a new dome-detection algorithm, insensitive to background and foreground noise, which detects probes of any intensity. We detail our two segmentation methods and our EM algorithm, and discuss the strengths of our techniques compared with state-of-the-art approaches. Both our segmentation methods are unsupervised, automatic, and require no training nor tuning: as a result, they are directly applicable to a wide range of medical images. We have used them as part of a large-scale project for the improvement of prenatal diagnostic of genetic diseases, and tested them on more than 2,100 images with nearly 14,000 nuclei. We report 99.3% accuracy for each of our segmentation methods, with a robustness to different laboratory conditions unreported before.

1 Introduction

Over the past twenty years, the age of pregnancy has been rising significantly, with increased risk of genetic disease for the children. Thanks to the progress made by research in genetics, many genetic diseases can now be treated at birth, sometimes even during pregnancy. However, current diagnostic methods require invasive procedures such as amniocentesis or cordocentesis, increasing the risk of miscarriage. It is known that a few fetal cells enter the maternal circulation: this opens the promise of a non-invasive diagnostic alternative. Isolating these cells non-destructively would give access to the whole genetic material of the foetus. Yet, such cells are rare: a sample of maternal blood will contain roughly 1 in 10^6 fetal cells, a ratio that can be reduced to about 1 in 10^4 with enrichment methods [1]. Computer vision can make their detection easier.

The cells used in this work are leucocytes. Their nuclei are treated with a blue fluorescent marker, and their telomeres with green fluorescent probes (see

A. Leonardis, H. Bischof, and A. Pinz (Eds.): ECCV 2006, Part IV, LNCS 3954, pp. 437–450, 2006.

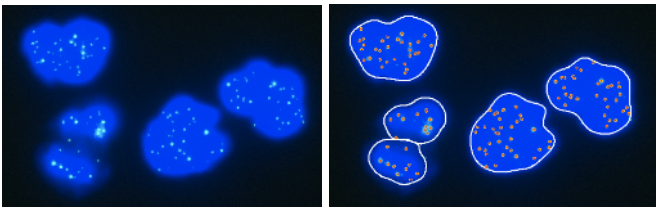


Fig. 1. Left: Image showing leucocyte nuclei (in blue), containing probes attached to the telomeres (in green). Right: nuclei and probes segmented with our method.

Fig. 1). Telomeres are the ending parts of chromosomes, and more abundant in fetal nuclei than in maternal ones. Measuring the green fluorescence in each nucleus is expected to be sufficient to single out the rare fetal nuclei within a sample of maternal blood, but it is still open to cytology research to assert it. The need for automatic image processing in this field of medical research is critical [2], and our work aims to meet this need.

Two significant issues impact the segmentation of such images. First, the nucleic fluorescence spreads into the background region immediately surrounding. Fig. 2 shows a typical nucleus, a profile of intensities across the image, and the histogram of the three segmented regions: it appears that the background is greatly affected near the nucleus. We use the term *illuminated background* to denote the background region where the intensity is increased by a nearby nucleus. Its extent is delimitated with a dotted line in Fig. 2. This region, hardly noticeable by eye, is critical to the correctness of the segmentation of the nuclei. The other issue is what we call *foreground noise*: unattached probes that cannot be perfectly washed out of the preparation sometimes accumulate as clumps and appear as bright spots, similar in intensity and shape to actual probes.

Previous work in this field is abundant, but is not usable in our context, where a large number of images taken in various laboratory conditions has to be analysed with minimal user interaction, and where, to reduce the time needed for diagnosis, the nuclei used are not cultured – as a result, they do not appear as convex and smooth as most image processing methods for nuclei segmentation require. In the following paragraphs we review the state-of-the-art methods. First, we review nucleic segmentation, then probe segmentation, and finally complete systems that are used in laboratories for similar purposes.

Nucleic segmentation methods can be classified in four categories: background subtraction, thresholding, watershed, and energy-based methods. For background subtraction, the background is generally either considered as uniformly noisy (with a histogram consisting of one Gaussian curve) [3], or is modeled with a reference image containing no objects, taken from an empty slide [2]. However, none of these models is satisfactory because of the background illumination near the nuclei. Threshold-based methods commonly used in cytometry, such as Otsu's [4] or Kittler and Illingworth's [5], assume that histograms are bimodal, or even consist of two Gaussians: this would be a crude and unrealistic estimate for our images (see histogram of a typical image in Fig. 2). Global thresholding

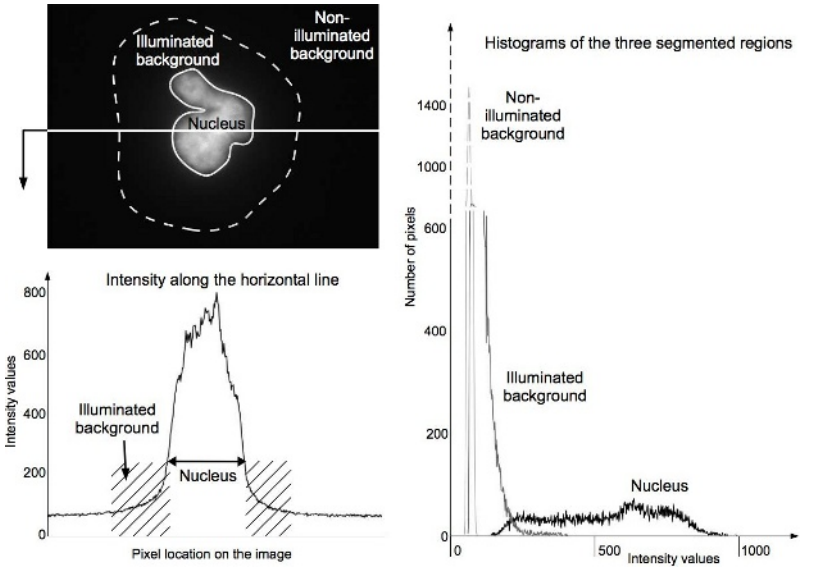


Fig. 2. Background illumination. Top left: image of a nucleus. Bottom left: intensity values along the horizontal line across the image. Right: Histograms of the three segmented regions (The vertical scale is bilinear, as reflected by the values on the side).

is bound to fail, because of cross-image intensity variations. Regarding local thresholding, the construction of a threshold surface introduces extra size and smoothness parameters, which have to be tuned, and are sensitive to the size and number of objects in an image. Watershed-based methods [6] are notorious for oversegmenting images; they require pre-processing with morphological operations to smooth the image, and post-processing to merge contiguous regions using shape, size and texture criteria. Most morphological operations use a filter, whose size and profile are to be tuned according to the image's properties and the smoothness required, and are thus little robust when automated. Region-merging is a long process, where the criteria for merging depend on the watershed results, and have to be tuned as well. Finally, energy-based methods, such as active contours [7], level sets, or graph cuts, require initialisation, internal energies modeling the final shape, external energies modeling the borders' characteristics, and parameters to balance them: these are difficult to tune even manually. Furthermore, in our context, the various nucleic textures and their impacts on the surrounding background are hard to model as local energy terms. Besides, the irregular shapes of uncultured nuclei elude typical internal energy terms. To summarise, these methods model the objects' characteristics independently, and require a complex parametrisation to link these models together.

Regarding probe segmentation, most methods are designed to segment only large bright probes, and usually less than four per nucleus. Existing probe-finding methods filter the image and threshold the intensities in order to keep a given percentage of bright pixels [8]. Two significant problems arise from the

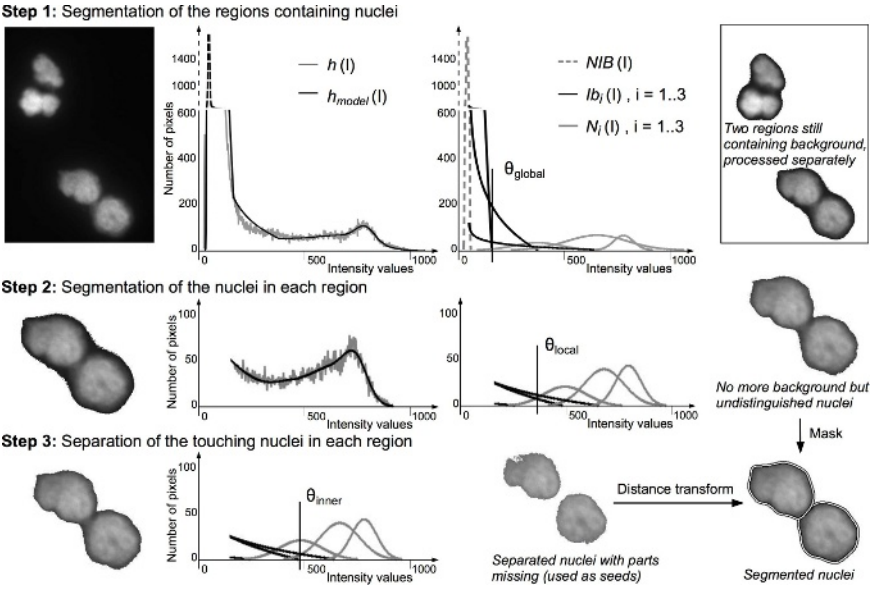


Fig. 3. The steps of nucleic segmentation. Top row, left to right: original image; its histogram with our adjusted model superimposed; the components of our model, used to find a global threshold removing most of the background; the segmented image, with two segmented regions containing nuclei but also some background. Middle row: one of these regions; its histogram and our adjusted model; the components of our model, used to find a local threshold, removing the remaining background; the segmented region, with no more background, but where the two nuclei are still undistinguised: it will be used as a mask for the final step of the segmentation. Bottow row: the segmented region; the components of our model, used to find an inner threshold, removing the darkest parts of the nuclei; the segmented nuclei, isolated but with parts missing: they are used as seeds, and grown with a fast distance transform within the mask defined above; the resulting segmentation of the region, with two separated nuclei and no background. See Section 3 for more details.

foreground noise. First, it flaws any histogram-based method by significantly increasing the number of high-intensity pixels. Second, as it is segmented as probes, existing methods need post-processing with a carefully designed classifier to distinguish it from the actual probes [9]. Also, existing systems measuring probes intensity require calibration, usually using a set of fluorescent beads [8], and are sensitive to changes in the fluorescence of the markers over time.

Finally, there are several integrated systems that are used in laboratory conditions for similar purposes; however, none of them is either automatic enough or general enough for clinical application. Many systems require expert human intervention at some point during the segmentation of each nucleus [10, 11]. Automatic systems are not as general-purpose as ours: [12] require an extra specific marker on the nuclei's borders, while [13] only segments isolated convex elliptic nuclei; systems such as Castleman's [2] or Netten's [14], are only applicable to images with few nuclei and few probes.

Our approach to segmentation is designed to avoid from the beginning the problems mentioned above. For the nuclei, we use a unified model which encompasses seamlessly the background noise, the nucleic textures and the nuclei's alterations to the background. The parameters of our model are intuitive and are automatically adjusted to every image, using an adaptation of an Expectation-Maximisation algorithm. This way, our model adapts to images of varying intensities and qualities, with no prior assumptions, training or manual tuning. We use our model to find successive threshold values, first global then local, and to isolate touching nuclei – one of the most difficult tasks in cytometry. The three steps of our nuclei segmentation are illustrated in Fig. 3. For probe segmentation, we use a new dome-detection algorithm which is insensitive to background and foreground noise, and detects any number of probes of any intensity, with no calibration required. After segmentation, the locations and measures of the nuclei are stored in an XML database for later retrieval.

This article is organised as follows: in Section 2, we detail our novel model for histograms. In Section 3, we describe our Expectation-Maximisation algorithm adapted for histogram modeling. In Section 4, we present our new dome-detection method applied to probe segmentation. In Section 5, we compare our method with a typical watershed-based segmentation, discuss the results, and present the results obtained with our software to compare individuals' ages using telomeres intensities – a critical issue for non-invasive prenatal diagnosis as mentioned earlier. We conclude in Section 6 with an overview of our future work.

2 Model of the Histogram of an Image

As illustrated in Fig. 2, the histogram of a typical image consists of three overlapping parts: a sharp peak in the lowest values, a sharply decreasing curve in the medium values, and a plateau in the highest values. They correspond respectively to the non-illuminated background (NIB), illuminated background (IB), and nuclei (N). In this section we present the parametric functions we use to model each part, and emphasis our new model for the illuminated background.

Let $h(I)$ be an image histogram, consisting of parts NIB, IB and N, which we model with $h_{model}(I)$. We assume NIB contains A_b pixels, has a mean value I_b, and is affected by Gaussian noise of standard deviation σ_b. It is modeled with:

$$NIB(I) = \frac{A_b}{\sqrt{2\pi}\sigma_b} \exp\left(-\frac{(I - I_b)^2}{2\sigma_b^2}\right) . \tag{1}$$

The part of the histogram corresponding to the highest intensities, N, reflects the nuclei's textures. They are very variable, within and across samples: in particular, variations affect the range of intensity values, the shape of the histogram and the number of peaks in it. Also, saturation can occur at high intensities, depending on the hardware used for imaging. To overcome these problems we model the nuclei's histograms with sums of Gaussians: this is both robust and flexible enough for our needs. As we do not know in advance how many nuclei an image contains, nor how many Gaussians are needed for each texture, we introduce a

new parameter, n, the number of Gaussians modeling the nuclei. Each of these Gaussians i will model A_i pixels, with mean I_i and deviation σ_i:

$$N_i(I) = \frac{A_i}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(I-I_i)^2}{2\sigma_i^2}\right) \ , \ 1 \leq i \leq n \ . \tag{2}$$

Next we detail the model we use for the illuminated background. Let I_0 be the intensity at the nucleus' border, at distance R_0 from its center, and let I_b be the mean intensity of the non-illuminated background (see Eq.(1)). We model the intensity in the illuminated background, along a line normal to the nucleus' border, with a decreasing exponential (see Fig. 4):

$$I(r) = I_b + (I_0 - I_b) \cdot \exp\left(-\frac{r-R_0}{\rho}\right), \text{ for } r \geq R_0 \ , \tag{3}$$

where ρ is a constant controlling the slope of the intensity decay. This model cannot be fitted directly to an image for segmentation purposes, as it requires a prior segmentation of the nuclei. Nevertheless, it can be used to derive a model of the illuminated background's histogram. This latter model can be adjusted to the image histogram, as detailed in the next section. In the remaining of this section, we explain how we derive that model.

The expression of $I(r)$ in Eq. (3) can be inverted to define $r(I)$. This can be used to express the number of points $dn(r) = 2\pi\, r\, dr$ at distance r from the nucleus, as a function of the intensity, $dn(I)$. By integrating $dn(I)$ between I and $I+1$, we obtain – by definition – the illuminated background's histogram. Introducing the new parameters $\alpha = \frac{\rho}{R_0}$ and $A = \pi R_0^2$, we obtain:

$$IB(I) = 2\, A\, \alpha \int_I^{I+1} \left(1 - \alpha \ln \frac{I-I_b}{I_0-I_b}\right) \frac{dI}{I-I_b} \ . \tag{4}$$

Eq. (4) is independent of the nucleus' actual shape: it only depends on its area A and the dimensionless parameter α, controlling the extent of the illumination relative to the nucleus' size. It can be easily integrated with the change of variables $X = \frac{I-I_b}{I_0-I_b}$. Also, this model is to be fitted to the histogram at values above the mean background value, with $I - I_b \gg 1$. Thus, a first-order expansion of $IB(I)$ with respect to $\frac{1}{I-I_b}$ is enough for our purpose. This leads to the definition of our new model for the histogram of the illuminated background:

$$IB(I) = \frac{2\, A\, \alpha^2}{I-I_b} \ln\left(\frac{I_0-I_b}{I-I_b}\right) \ . \tag{5}$$

The expression of $IB(I)$ in Eq. (5) is illustrated in Fig. 4, to the right. Its two parameters A and α correspond respectively to the area of the nucleus creating the illumination, and to the spatial decay of the illumination.

To link this model with that of the nuclei, we assume that each Gaussian modeling the nuclei's textures creates part of the background illumination. Let $IB_i(I)$ model the background illuminated by $N_i(I)$. Three of its four parameters are constrained by the rest of the model: namely, the area causing the

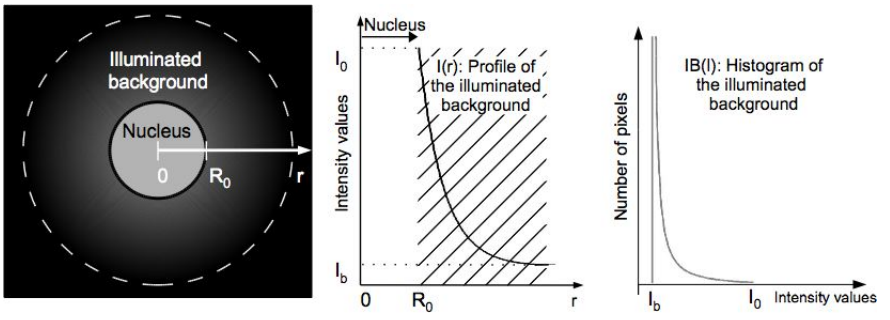


Fig. 4. Illustration of our model. Left: model of a circular nucleus. Center: model of the intensity values in the illuminated background, $I(r)$, defined in Eq. (3). Right: model of the histogram of the illuminated background, $IB(I)$, defined in Eq. (5).

illumination $A = A_i$ (see Eq. (2)), the mean intensity I_b of the non-illuminated background (see Eq. (1)), and the intensity I_0 at the nucleus' border. We set it to $I_0 = I_i - 2\sigma_i$. Using Eq. (2), it corresponds to the darker 5% of the nucleus' pixels, which we consider to be on the boundary. This way, the functions modeling the illuminated background are:

$$IB_i(I) = \frac{2\ A_i\ \alpha_i^2}{I - I_b}\ \ln\left(\frac{I_i - 2\sigma_i - I_b}{I - I_b}\right)\ ,\ 1 \leq i \leq n\ . \tag{6}$$

Let Φ_n be the set of all the functions used, which are defined in Eqs. (1), (2) and (6). The model of the histogram is:

$$h_{model}(I) = \sum_{g \in \Phi_n} g(I)\ ,\ \text{where}\ \Phi_n = \{NIB\} \cup \bigcup_{1 \leq i \leq n} \{IB_i, N_i\}\ . \tag{7}$$

It depends on the $4(n+1)$ parameters $n, A_b, I_b, \sigma_b, \{A_i, I_i, \sigma_i, \alpha_i\}_{1 \leq i \leq n}$. They are adjusted to an image histogram using an Expectation-Maximization algorithm, as detailed in the next section.

3 Expectation-Maximisation Algorithm for Histogram Modeling

Let $h(I)$ be an image histogram, consisting of parts NIB, IB and N, which we model with $h_{model}(I)$. Each intensity I in the histogram contains a proportion of pixels modeled by each function of our model. We define this proportion as:

$$p_f(I) = \frac{f(I)}{\sum_{g \in \Phi_n} g(I)}\ ,\ \forall f \in \Phi_n = \{NIB\} \cup \bigcup_{1 \leq i \leq n} \{IB_i, N_i\}\ . \tag{8}$$

Knowing these proportions is enough to define the successive thresholds needed for our nucleic segmentation, as detailed at the end of this section. Given all

Algorithm 1. Expectation-Maximisation algorithm for histogram modeling

1: **for** $n = 1$ to 6 **do**
2: *Initial E-step:* set the initial proportions as in Table 1
3: **for** 100 times **do**
4: *M-step:* compute the parameters with Eqs. (9) and (10).
5: *E-step:* update the proportions with Eqs. (1), (2), (6), (8).
6: *Evaluation:* measure the error between the model and the histogram as $error = \sum_I \left(1 - \frac{h_{model}(I)}{h(I)}\right)^2$. Store the model if the error is the lowest.
7: **end for**
8: **end for**
9: **return** the model with the lowest error.

the parameters of the model, these proportions can be computed using Eqs. (1), (2), (6) and (8). Reciprocally, given all the proportions $p_f(I)$, $\forall I$, $\forall f \in \Phi_n$, the parameters of the model can be computed as described below. However, neither the parameters nor the proportions are available in the first place. This type of problem is commonly solved by Expectation Maximisation [15]. The EM algorithms commonly used in computer vision are adapted to mixture of Gaussian models. The algorithm we present as Algorithm 1 is adapted to histograms: the steps are the same, only the equations are different.

We now explain how to compute our model's parameters given a histogram $h(I)$ and all the proportions $p_f(I)$. The parameters of the Gaussians *NIB* and N_i are computed as the total, mean and deviation of a weighted histogram [16]:

$$A_i = \sum_I p_{N_i}(I)\, h(I)\,;\; I_i = \sum_I p_{N_i}(I)\, h(I)\, I\,;\; \sigma_i^2 = \sum_I p_{N_i}(I)\, h(I)\, (I - I_i)^2\,. \quad (9)$$

The only unconstrained parameters of the functions IB_i are computed as:

$$\alpha_i = \frac{1}{|\ln \epsilon|} \sqrt{\frac{1}{A_i} \sum_{I=I_\epsilon}^{I_i - 2\sigma_i} p_{IB_i}\, h(I)}\,, \text{ where } \varepsilon = 0.1\,, I_\varepsilon = I_b + \varepsilon\,(I_0 - I_b)\,. \quad (10)$$

(See Appendix for details). The complete segmentation of the nuclei is performed in three steps, as illustrated on Fig. 3. First, we use the algorithm above to adjust our parametric model to the image histogram. Let:

$$\theta_{global} = \max\{I, \forall i, \min(p_{NIB}(I), p_{IB_i}(I)) \geq p_{N_i}(I)\}\,.$$

Below θ_{global}, all intensities contain more points from *NIB* or IB_i than from the corresponding N_i, and it is the highest such value. This is the global threshold we use to discard the non-illuminated and part of the illuminated background. Then, in each of the segmented regions, we apply the same algorithm to find a model of the histogram (without the *NIB* function this time). In the same way, we find the highest intensity containing more points from IB_i than from N_i, and use it as a local threshold θ_{local}. The newly segmented components

Table 1. Initial proportions' values for the EM algorithm. The histogram range $[0, I_{max}]$ is divided in $n+2$ parts. Intuitively, most dark pixels are modeled by *NIB*, most bright pixels by either one of the N_i, and most of the remaining pixels by the IB_i (see Fig. 2).

Intensity	$0 \ldots \frac{1}{n+2} I_{max}$	$\frac{1}{n+2} I_{max} \ldots \frac{2}{n+2} I_{max}$	$\frac{j}{n+2} I_{max} \ldots \frac{j+1}{n+2} I_{max}$ for $2 \leq j \leq n+1$
$p_{NIB}(I)$	0.9	0.1	0
$p_{IB_i}(I)$	$0.1/n$	$0.9/n$	$0.1/n$
$p_{N_i}(I)$	0	0	$0.9/n$ if $i = j-1$, 0 else

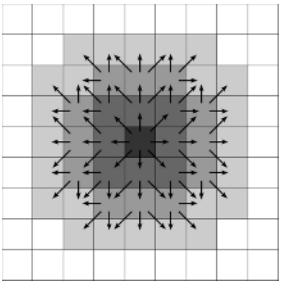
do not contain background anymore; however, they might contain more than one nucleus. This problem is often solved by splitting components into convex parts [17], but cannot be applied here – uncultured nuclei may be concave. Instead, we consider the components' textures, which are already modeled by the N_i. We assume that there are several distinguishable nuclei in a component if it contains dark paths separating several bright parts. The threshold we use to define dark and bright for this test, called inner threshold θ_{inner}, is the lowest of the I_i. The connected components above θ_{inner} are considered as seeds: each one marks a unique nucleus. Then we extend the seeds into the regions above θ_{local} using a fast distance transform [18], and obtain the segmented nuclei. Fig. 3 shows the three steps of the segmentation, with the models adjusted to the histograms. Another example of a segmented image is shown in Fig. 1.

4 Dome-Finding Algorithm for Probe Segmentation

Once the nuclei are segmented, their telomere contents are to be evaluated, by segmenting the fluorescent probes in the green channel. Probes appear as small spots, each about a dozen pixels big. Background illumination is observed around the probes as well; however we cannot apply the same segmentation method as for the nuclei. This is for practical reasons: on a typical image, nuclei represent about 8% of the pixels in the image, and the illuminated background about 50%; but the probes only represent 0.3%, and the background around them, 2%. Adjusting our model using so few pixels would not be reliable enough.

Our novel method to segment probes is based on the following observations. Both background and foreground noise are characterized by high densities of local intensity maxima, distant by two or three pixels. Conversely, probes correspond to local maxima surrounded by pixels of decreasing intensity and few, if any, other local maxima within a distance of two or three pixels. Thus, we developed a peak-detection method sensitive to the density of local maxima. In addition, as it only measures pixels intensities relatively to their neighbours, our method can detect probes of high and low intensities, unlike traditional probe-finders restricted to few bright probes [2].

We segment probes as domes, starting from local maxima and gradually including neighbours if they form a dome around them. If a dome is large enough, we mark it as a probe; otherwise, we reject it. Around each local maximum, we consider three sets of neighbouring pixels, at increasing distances, as illustrated by different shades on gray in the left of Fig. 5. They form the level



140	132	145	155	148	163	146	141	127
152	147	179	182	195	177	139	140	142
134	168	197	256	281	230	188	143	131
168	186	242	333	359	347	192	142	157
160	188	270	342	442	336	222	163	143
155	208	245	319	349	257	185	159	144
183	181	245	244	234	192	171	142	130
148	157	176	176	152	142	145	134	138
146	139	144	133	140	140	137	127	126

1034	1057	1003	1023	1050	961	1124	1050	997
1050	978	1078	978	949	1064	1050	1064	1036
943	1050	1050	1086	1043	1064	991	1078	983
1108	1057	1071	1078	1003	1009	1050	1023	893
1093	1086	1064	1086	1009	1050	1029	1041	889
1029	1023	1043	1071	972	1078	984	925	810
1093	997	1064	997	1043	955	1043	908	807
997	1029	1016	991	1064	1050	938	818	707
1043	1003	1043	1050	1029	1064	894	814	706

Fig. 5. Left: the three level sets around a pixel, in three shades of gray, and all the downhill neighbours, indicated by the arrows. Middle: actual pixels values in the neighbourhood of a probe, and the segmented dome in gray. Right: actual pixels values in a zone of foreground noise; none of the local maxima is surrounded by a proper dome.

sets[1] of the dome, and approximate the shape of the probes. Formally, let p_M be a local maximum. The first level set consists of the 8 closest pixels: $LS_1 = \{p, d_\infty(p, p_M) = 1\}$. In the second level set are the twelve closest pixels that are not already in LS_1: $LS_2 = \{p, p \notin LS_1 \wedge d_1(p, LS_1) = 1\}$. Each pixel in these sets is assigned three neighbours, as indicated by the arrows in Fig. 5: we refer to them as *downhill neighbours*. If a pixel has a higher intensity than all its downhill neighbours, it is marked as being part of the dome.

By design, one complete dome corresponds to one probe. However, it happens that two probes are very close (two pixels apart), or that a probe is bigger than a dome (and contains two local maxima). In both cases, the dome construction above leads to two domes having one side in common that does not meet the downhill constraint, and which is therefore not included in any of the two domes. Since these cases are at the borderline but still valid, we accept domes with up to one of their four sides missing. Formally, a dome is marked as a segmented probe if it contains at least 75% of the pixels in LS_1 and 75% of LS_2. Domes with more pixels missing are rejected as noise (see middle and right of Fig. 5).

5 Results and Discussion

We start this section by presenting quantitative results obtained with our novel segmentation methods. Our data set contains 2,166 images, with nearly 14,000 nuclei and 317,000 probes overall. We compare our results with a typical watershed-based nucleic segmentation method. We also present an application of our methods, to compare the telomere intensities of two different individuals.

5.1 Accuracy of the Segmentation Methods

Nucleic Segmentation. We implemented our method on an iMac with a 1.8GHz PowerPC G5 and 1Gb of RAM, and processed our full dataset. For comparison,

[1] Here, the term *level set* refers to its original definition in topology, not to the segmentation method with the same name.

Table 2. Results of the nuclei segmentation

Method	*Number of nuclei*	*Correctly segmented*	*Over segmented*	*Under segmented*	*Missed*	*Non existing*
Watershed	2,779	2,232 - 80.3%	211 - 7.6%	212 - 7.6%	60 - 2.2%	64 - 2.3%
Ours	13,917	13,823 - 99.3%	31 - 0.22%	50 - 0.36%	8 - 0.06%	5 - 0.03%

we also implemented the watershed-based segmentation method for cytometry described in [19], as follows. The image is thresholded globally with the value found by the isodata algorithm; a distance transform is applied to the resulting image, followed by an h-dome extraction; the domes extracted are used as starting-points for the watershed algorithm, applied to the gradient transform of the original image. We implemented this method under the same conditions as ours, and tested it on a sample of 800 images, containing over 2,000 nuclei. The results of these two segmentation methods are listed in Table 2, and discussed below.

In terms of runtime, our method segments one nucleus in about one second (with no particular programming optimisation), which is about three times faster than the watershed-based one, and several times faster than manual segmentation. Our method has linear complexity, and no particular memory requirements (a histogram and the segmentation results). Once the histogram is built, our EM algorithm runs in constant time; the thresholding steps require one image scan. Conversely, the watershed-based method needs to store extra intermediate images, and the h-dome extraction requires an unknown number of image scans. Besides, this method takes several minutes to process images with no nuclei, and systematically segments objects in them. Our method processes empty images correctly and faster than images with nuclei.

The quantitative results, shown in Table 2, are significantly better with our method, which is due to its two main features. First, it finds the nucleic borders using a succession of threshold values that are adapted for each part of the image containing nuclei, while the other method uses a single global threshold. As a result, many more nuclei are missed, when darker that the global threshold, and many non-existing nuclei are segmented, which are in fact bright background regions. Secondly, our method uses a texture model to separate touching nuclei, and gets very low oversegmentation (when a nucleus is segmented in more than one object) and undersegmentation (when more than one nucleus are segmented as one object). Conversely, to find seeds, the watershed-based method replaces the nuclei's textures with a distance transform, which amounts to using only the nuclei's borders to separate them. This approach is bound to fail with uncultured nuclei, having concavities, as illustrated by the higher over- and undersegmentation rates. Similar quantitative results are reported in [20], for the same watershed-based method and for a contour-based method. Both methods correctly segment 80% of similar nuclei, and over- and under-segment a total of 15% of the dataset. Using successive watershed-based and contour-based methods, [13] reports a 99.4% segmentation accuracy, but their method requires the prior rejection of all the non-isolated, non-elliptic nuclei – corresponding to 30% of their data, but more than 50% of ours.

Our results can still be improved by basic post-processing. Missed objects cannot be recovered, but are very rare in the first place, while non-existing nuclei are hardly an issue. Oversegmentation happens when θ_{inner} is too high, and can be detected by the small size of the parts oversegmented. Undersegmentation is due to either θ_{global} or θ_{local} being too low, and results in bigger than average objects, which can be detected as such and processed with higher thresholds.

Probe Segmentation. The aim of segmenting probes is to measure the total fluorescence inside a nucleus, so oversegmentation is not an issue. Undersegmentation is prevented with our method, as a segmented probe has a minimum dome size. Overall, 99.3% of the probes were correctly found with our method. About 0.3% were missed, too wide to be detected as one dome. Most of them were dark, with little effect on the total fluorescence measured inside the nucleus; very rare wide and bright probes were ruled out as foreground noise (less than 0.1%). Finally, about 0.4% of the segmented objects were background, not probes; they were dark and did not affect the final measures.

5.2 Comparison of Individuals Ages Using Telomeres Intensities

As an application, we used our method to quantify the intensity of probes appearing in the nuclei for two individuals. This test was conducted to assess the differences in telomeric intensities between individuals of different ages. In particular the two populations of nuclei were not mixed. After using our segmentation, our program rejected the nuclei which were cropped at the edges of images. The results are shown in Fig. 6. The first histogram shows that the same number of probes per nucleus were segmented for the two individuals; the second histogram shows that the probes in the fetal nuclei are brighter. These result show that our method does not introduce bias in the number of probe segmented, and that there is a promising distinction between the individuals.

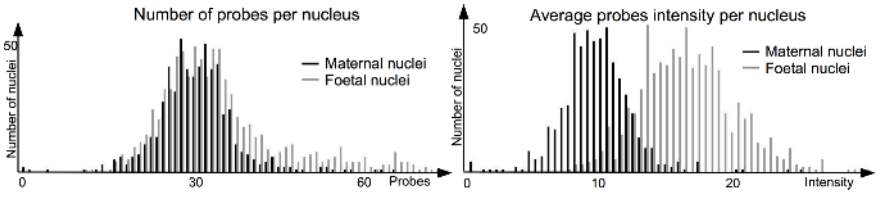


Fig. 6. Quantitative measures performed using our method on two populations of nuclei

Our software has proven reliable and robust enough to produce these results. Reducing the overlap between the two histograms is a subject for cytology research. Here again computer vision may help, as detailed in the next section.

6 Conclusions and Future Work

In this article we have detailed our new segmentation methods, presented a quantitative comparison of our nucleic segmentation with the widely used watershed

method, and shown an application of our software for medical research. The nucleic segmentation method we developed is based on a new model of the image histogram, achieves a 99.3% accuracy and, to the best of our knowledge, is more robust and automatic than previously published work on this field. Also, we have presented some ideas to improve this rate further. As for the telomere probes, our method is robust against all sources of noise, and is also 99.3% accurate.

The segmentation techniques we developed can be used for various cytometric tasks. We have used it along with our telomeric segmentation method for a project of improved diagnostic methods. The quantitative results we have obtained show a promising distinction between the telomere intensities in individuals of different age. To improve the difference and reach the stage where a fetal nucleus can be detected within a population of maternal nuclei, we are participating in further work with cytologists. Not all uncultured nuclei will be usable in the final stage, where their genetic content is investigated. Some are damaged during the early processing of sample blood, and could be rejected before measuring the telomere intensities. These unusable nuclei can be detected by an expert cytometrist by their shapes. We are currently working with such experts on an automatic shape analysis of the segmented nuclei: our early work includes measuring the nuclei's concavities and using low-order Fourier reconstructions to define usability criteria. Rejecting these unusable nuclei before segmenting the telomeres would make our final comparison of populations more conclusive. At that stage, we will be in a stronger position to tell if this approach to non-invasive diagnostic alternative is reliable enough for a future clinical application.

The author thanks Prof. Clocksin, Dr Bray and Dr McCollum for their help and Prof. Hulten and Dr Ariosa for providing the images.

References

1. Hohmann, H., Michel, S., Reiber, W., Gunther, M., Claussen, U., von Eggeling, F.: How to enrich and analyse fetal cells from maternal blood. In: Fetal Cells and Fetal DNA in Maternal Blood, Karger (2001) 47–55
2. Merchant, F.A., Castleman, K.R.: Strategies for automated fetal cell screening. Human Reproduction Update **8** (2002) 509–521
3. Fang, B., Hsu, W., Lee, M.L.: On the accurate counting of tumor cells. IEEE Transactions on Nanobioscience **2** (2003) 94–103
4. Tanaka, T., Murase, Y., Oka, T.: Classification of skin tumors based on shape features of nuclei. In: Medicine and Biology Society. Volume 3. (2002) 1064–1066
5. Wu, K., Gauthier, D., Levine, M.D.: Live cell image segmentation. IEEE Transactions on Biomedical Engineering **42** (1995) 1–12
6. Lin, G., Adiga, U., Olson, K., Guzowski, J., Barnes, C., Roysam, B.: A hybrid 3D watershed algorithm incorporating gradient cues and object models for automatic segmentation of nuclei in confocal image stacks. Cytometry **56A** (2003) 23–36
7. Lehmann, T., Bredno, J., Spitzer, K.: On the design of active contours for medical image segmentation. Methods of Information in Medicine **42** (2003) 89–98
8. Poon, S.S., Martens, U.M., Ward, R.K., Lansdorp, P.M.: Telomere length measurements using digital fluorescence microscopy. Cytometry **36** (1999) 267–278
9. Clocksin, W.F., Lerner, B.: Automatic analysis of fluorescence in-situ hybridisation images. In: British Machine Vision Conference. (2000) 666–674

10. Pelikan, D.M.V., Mesker, W.E., Scherjon, S.A., Kanhai, H.H.H., Tanke, H.J.: Improvement of the Kleihauer-Betke test by automated detection of fetal erythrocytes in maternal blood. Cytometry Part B (Clinical Cytometry) **54B** (2003) 1–9
11. de Solorzano, C.O., Garcia Rodriguez, E., Johnes, A., Pinkel, D., Gray, J., Sudar, D., Lockett, S.: Segmentation of confocal microscope images of cell nuclei in thick tissue sections. Journal of Microscopy **193** (1999) 212–226
12. de Solorzano, C.O., Malladi, R., Lelievre, S., Lockett, S.: Segmentation of nuclei and cells using membrane related protein markers. Microscopy **201** (2001) 404–415
13. Bamford, P., Lovell, B.: Method for accurate unsupervised cell nucleus segmentation. In: IEEE Engineering in Medicine and Biology. Volume 1. (2001) 133–135
14. Netten, H., Young, I., van Vliet, L., Tanke, H., Vroljik, H., Sloos, W.: Automation of fluorescent dot counting in interphase cell nuclei. Cytometry **28** (1997) 1–10
15. Theodoridis, S., Koutroumbas, K.: Pattern Recognition. 2^{nd} edn. Elsevier (2003)
16. Smith, S.W. In: The scientist and engineer's guide to digital signal processing. California Technical Publishing (1997) 11–34
17. Kutalik, Z., Razaz, M., Baranyi, J.: Automated spatial and temporal image analysis of bacterial cell growth. In: BMVA Spatiotemporal Image Processing. (2004)
18. Felzenszwalb, P.F., Huttenlocher, D.P.: Distance transforms of sampled functions. Technical report, Cornell University (2004) TR2004-1963.
19. Malpica, N., de Solorzano, C.O., Vaquero, J.J., Santos, A., Vallcorba, I., Garcia-Sagrado, J.M., del Pozo, F.: Applying watershed algorithms to the segmentation of clustered nuclei. Cytometry **28** (1997) 289–297
20. Restif, C., Clocksin, W.: Comparison of segmentation methods for cytometric assay. In: Medical Image Understanding and Analysis. (2004) 153–156

Appendix

We use the notations of Section 2. The sum of histogram values of the illuminated background between any two values I_1 and I_2 is: $\sum_{I=I_1}^{I_2} IB(I) = \int_{I_1}^{I_2+1} dn(I)$. However, $dn(I)$ is only defined and positive between I_b and I_0, and is not summable near I_b. Let $I_\varepsilon = I_b + \varepsilon\,(I_0 - I_b)$, where $\varepsilon \in (0,1)$: $\sum_{I=I_\varepsilon}^{I_0-1} IB(I) = \int_{I_\varepsilon}^{I_0} dn(I)$. Developed to first order terms: $\sum_{I=I_\varepsilon}^{I_0-1} IB(I) = -2A\alpha^2 \int_{I_\varepsilon}^{I_0} \ln\left(\frac{I-I_b}{I_0-I_b}\right) \frac{dI}{I_0-I_b}$. Since $IB(I_0) = 0$, the sum can be extended to I_0, while the integral can be computed with a change of variable: $\sum_{I=I_\varepsilon}^{I_0} IB(I) = A\,\alpha^2 \ln^2 \varepsilon$. This gives the expression of α as a function of the histogram values: $\alpha = \frac{1}{|\ln \varepsilon|} \sqrt{\frac{1}{A} \sum_{I_\varepsilon}^{I_0} IB(I)}$.