An Adaptative-local and Geometric Approach for Automatic Chromosome Image Segmentation

Viridiana Rubí Calzada-Navarrete and César Torres-Huitzil
`vcalzada@tamps.cinvestav.mx, ctorres@tamps.cinvestav.mx`

Information Technology Laboratory
CINVESTAV-Tamaulipas, Mexico

Abstract. Karyotyping is a cytogenetic technique used to detect genetic aberrations. For many years, different computational techniques have been proposed to automate or assist in this task. One of the main stages in Karyotyping is image segmentation which consists of 3 main subtasks: object-background separation, object discrimination and union/overlap solver. This paper proposes a method towards an automatic chromosome image segmentation, specially for the two first subtasks. First a re-threshold process based on Sauvola local adaptive technique is applied to extract objects of interest from background. In the discrimination subtask, object geometric characteristics are used to discriminate between single and cluster of chromosomes. Experiments show encouraging results with 93% of precision in the re-threshold process and 93.5% in the chromosome discrimination subtask.

Keywords: Karyotyping, segmentation of cromosomas, re-threshold

1 Introduction

Chromosome karyotyping analysis is an important technique used to detect abnormalities and diagnosis By analyzing the number and shape of chromosomes found in an image, it is possible to detect abnormalities.

A chromosomes is comprised of two chromatid joined at a point called centromere and each chromatid turn consists of DNA strings that are positioned in different ways giving different structures and size variety for each chromosome. Chromosomes are commonly textured as ligth and dark strips called banding patterns.

Chromosomes are stained from cells of tissue, blood or amniotic fluid samples. Then a microscope is used to acquire an image of the cell division process.

An automatic karyotyping system involves a number of different steps: segmentation, feature extraction, and classification.

The segmentation stage resolves the problems that presents the image as: lighting problems that affect the extraction of the actual area of the chromosome which is caused by two factors, the amount of incident light from the source on the scene, such as interference with elements of capture, type of microscope the intensity of light from its lamp, leave stains on the background of the image that

© E. Castillo, J.C. Chimal, A. Uriarte, L. Cabrera.
Advances in Computer Science and Engineering
Research in Computing Science 58, 2012 pp. 3–16
Paper Recived 17-08-2012 and Acepted 07-11-2012

affect the outline of objects to perform the extraction and the other, the amount of light reflected by the stained chromosomes.

Low contrast, the band patterns comprising the chromosome and the centromere have different gray levels causing disadvantages in the identification of the actual shapes of the chromosome.

Union and overlapping chromosomes, is the most complex problem to solve because it depends on the position and number of chromosomes involved. The characteristics of each chromosome allow to identify and implement appropriate processes to separate and reconstruct them in the cases of union and overlapping.

For the lighting and low contrast problems, it is proposed to work with local adaptive thresholding techniques, allowing to find a threshold that adapts to changes in the gray level image according to the neighborhood of each pixel. For discrimination of single and clusters of chromosomes, a technique based on geometric and morphological features objects is proposed

The rest of the work is organized as follows. Section 2 describes the related work and justify the proposed idea. In section 3, "re-threshold" and discrimination techniques are presented. Section 4 shows results and finally in section 5, conclusions and future work are presented.

2 Related work

There are different proposed solutions to chromosome image segmentation; some of them need the intervention of an expert, however the research continues working on the improvement of techniques and results.

The first step in analyzing a chromosome image is the segmentation of individual chromosomes and chromosome clusters from the image background. Unfortunately, the high variability in chromosome and background fluorescence intensities makes the utilization of a global threshold inappropriate.

Some works like [1,2,8] suggest that to remove background objects is not sufficient to apply a global technique such as Otsu. Because of banding patterns and that centromere has a low contrast, chromosomes become confused with the background of the image resulting in poor extraction. Works like [4] use a global threshold, guided only with the most representative gray level in the image according to its histogram.

In [2,7], authors show that local adaptive techniques provide good results providing a technique that divides the original image into fixed windows and then apply Otsu thresholding in each of those windows. Then the threshold matrix is interpolated in order to reach the original image size. L.Ji [6] proposes a scheme based on a "re-threshold" with two global thresholds. Allowing a little more adjusted parameters in the second threshold.

For the discrimination of chromosomes L.Ji [6] proposes 3 features to discriminate objects including size, roundness and gray level. The drawback is that these feature values do not maintain a pattern, as they vary according to the sizes and gray levels between images.

Geometric features are considered for discrimination in [2,4] to separate single objects according to small or large sizes. While for the rest, skeleton endpoints are used to assign clusters of chromosomes for those objects with more than 2 endpoints, in other case they are singles.

3 Proposed solution

The method that is proposed in this paper includes the first two steps of the segmentation stage: extraction, whose aim is to separate from the background single and cluster of chromosomes, which are then passed on to the discrimination, where each object (either a single chromosome or cluster) is analyzed to test whether it is composed by a single chromosome or many of them, these latter chromosomes are separated from the rest because they need to be disentangled.

The proposed approach is depicted in figure 1. The first 2 steps of the segmentation are enclosed in boxes.

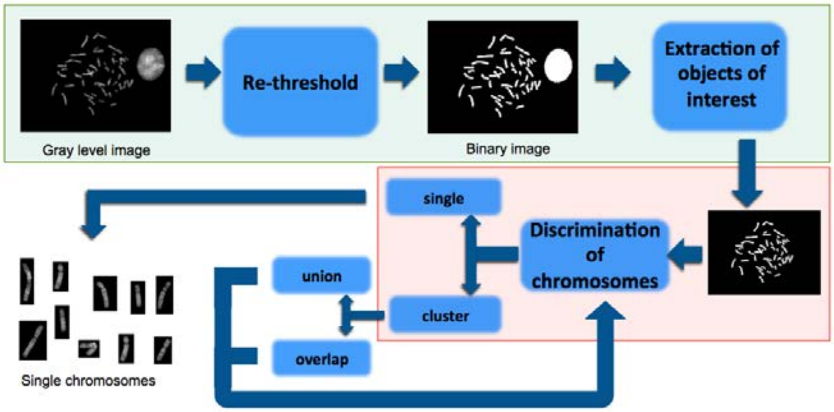


Fig. 1. Chromosomes segmentation process.

3.1 Extraction of objects of interest

According to the characteristics of the image, it is proposed to work with a local adaptive technique and not with a modified global technique such as Otsu. For that reason, the Sauvola's technique is proposed, which assigns a threshold to each pixel according to its neighborhood, returning a binary image highlighting the edges of objects in the image.

Sauvola's method is the result of an improvement to the method of Niblack, since it effectively reduces the effect of poor lighting (stain) in the image, besides being less sensitive to the parameter k which is a constant that ajusts the amount of pixels from the edge of the objects that are taken as part of it, and also appends a constant R which is the dynamic range of the standard deviation (see equation 1).

Sauvola's method adapts the threshold value according to the mean $m(x, y)$ and local standard deviation $\sigma(x, y)$, which are calculated in a window of size $b \times b$. Calculating the threshold for each pixel as:

$$T(x, y) = m(x, y) \cdot \left[1 + k \cdot \left(\frac{\sigma(x, y)}{R} - 1\right)\right] \tag{1}$$

where for binarization several authors determined that $k = 0.2$, $R = 128$ and $b = 15$ work for most of the cases.

The local adaptive threshold can properly identify chromosomes when they have a good contrast. But in general, the image needs a contrast enhancement to detect all the chromosomes. It is important to consider that the contrast enhancement highlights the chromosomes and partially solves lighting problems in the image.

Therefore it is proposed a Re-threshold, which is carried out with a double binarization to extract all the chromosomes without being affected by the background.

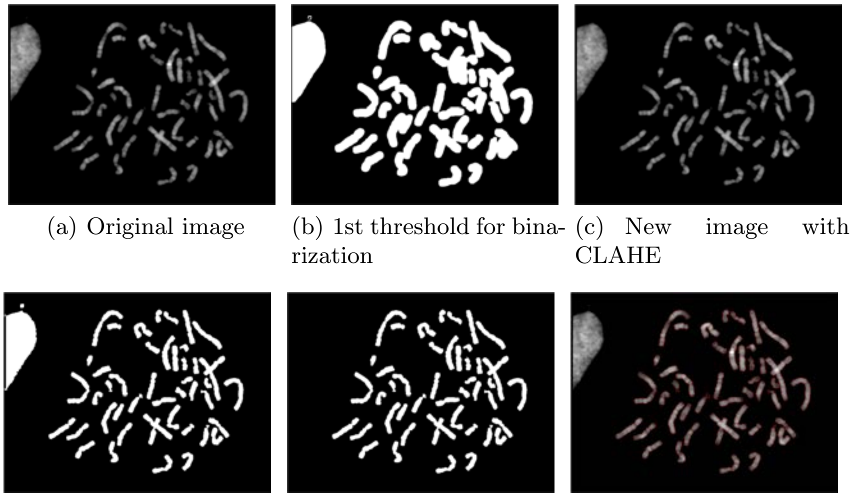
Re-threshold A first threshold is applied to the image in order to rescue areas that have a high probability of containing the objects of interest, thereby eliminating the background zones containing stains. Then the contrast on that area is enhanced with a CLAHE (Contrast Limited Adaptive Histogram Equalization) allowing to highlight the edges, bands and centromeres of chromosomes, and finally a new threshold to correctly extract the chromosomes is applied.

Thus, the proposed method for re-thresholding has the following steps:

1. The first threshold for the image $I(x, y)$ is obtained with the Sauvola technique and the parameters are those recommended in the literature except the window size that was chosen smaller for this case in orde to remove the area of interest without much detail, $k = 0.2$, $R = 128$ and $b = 9$.
2. The resulting image $BW1(x, y)$ is used as a mask to extract the interest region from the original image.
3. In the new image $I_m(x, y)$, a contrast enhancement with (CLAHE) is applied.
4. The second threshold is obtained with Sauvola technique and the parameters are those recommended in the literature, $k = 0.2$, $R = 128$ and $b = 15$.

Removal of unwanted objects At this stage the binary image $I1(x, y)$ of the previous step undergoes a procedure to identify and label each individual object. Therefore, each object can be selected and manipulated separately to remove undesirable objects with an erosion. If the object is not removed then it is a nuclei, otherwise the number of pixels and average gray level are evaluated. Then If the object is very small and dark, then it is eliminated.

In figure 2 is depicted how to perfom the extraction of the interest objects.



(a) Original image (b) 1st threshold for binarization (c) New image with CLAHE

(d) 2nd threshold for binarization in new image (e) Final image, only objects interest. (f) Edges of the final image on the original image

Fig. 2. Object extraction from background

3.2 Discrimination

Once interest objects have been extracted and labeled, the second step is analyze their features for discriminating between singles and clusters of chromosomes. The discrimination is performed with 4 geometric characteristics of chromosomes: area, solidity, eccentricity and skeleton endpoints.

The evaluation of the geometric characteristics discriminate chromosomes that are in a group. A threshold for each feature is estimated for each image, so that the threshold will depend on the number of predominate objects (individual or clusters) in the evaluation. The discrimination is performed with a decision tree using these characteristics. This indicates that after evaluating the threshold for each feature, a set of three images named individuals, clusters or suspects will be generated. The last image moves to the next assessment in the decision tree.

The discrimination stage receives labeled image as input. The evaluation sequence is displayed on figure 3

The first evaluation is performed with the solidity which is the ratio between the area of the object and its convex polygon[3].

$$S = \frac{Area}{Convex\_poligon\_area} \tag{2}$$

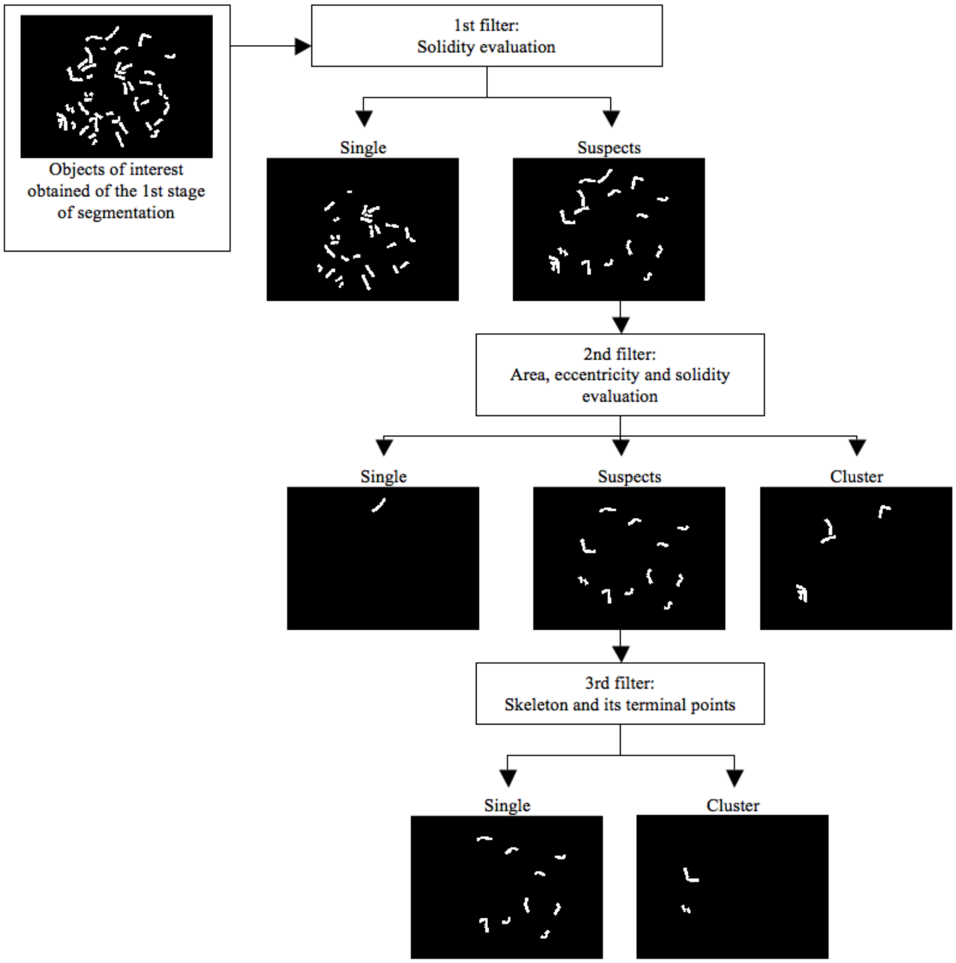


Fig. 3. Decision tree of chromosomes discrimination

where in 2: Area is the total number of pixels that form the object obj. For the binary image BW its area A is given by[5]:

$$A = \sum_{i=1}^{n} \sum_{j=1}^{m} BW[i,j] \tag{3}$$

and convex polygon area is the number of pixels that form the smallest convex polygon that can contain the selected object region[3].

Solidity S value is between 0 and 1. Objects with solidity close to 0 are those individual chromosomes whose form is completely straight, nearly straight or its size is small. While the rest are suspect chromosomes because still it is not possible to assign to a class. As there is no case where all chromosomes are overlapping, then the mean indicates the ratio between the number of group and individual chromosomes present in the image. Objects that are above the average of solidity are individual objects.

$$thS = \left(\frac{1}{n} \sum_{i=1}^{n} S_{obj_i} \right) - e \tag{4}$$

$$obj_i = \begin{cases} single & \text{if } S_{obj_i} \geq thS \\ suspect & \text{other case} \end{cases} \tag{5}$$

where e (equation 4) is a factor determined heristically to lower the threshold and rescue some individual chromosomes.

The second evaluation is the object area A and eccentricity E. It begins with the object area which also takes the mean of all objects in the image as a base to obtain the threshold. The mean indicates that if the object is very large (above the mean) is a group or very small (below the mean) is an individual object and small. Therefore, the thresholds for the area are as follows (see equations 6 - 8):

$$thA_s = 2.5\mu_A \tag{6}$$

$$thA_i = 0.5\mu_A \tag{7}$$

$$obj_i = \begin{cases} cluster & \text{if } A_{obj_i} > thA_s \\ single & \text{if } A_{obj_i} < thA_i \\ suspect & \text{other case} \end{cases} \tag{8}$$

For eccentricity which is the value obtained by dividing the minor axis between the major axis of the ellipse surrounding the object, it gives values between 0 and 1[3]. If the eccentricity is close to 1, it indicates that the object is a group of chromosomes while if it is close to zero corresponds to a single chromosome. There are cases in which the chromosome is curved in a "C" or "S" and could be confused with a group when it is not, for that reason there are two thresholds, one higher than the average eccentricity and very close to 1 which confirms that it is a group and the other below the average and very close to 0 which ensures that it is a single well are the thresholds (see equations 9 - 10):

$$thE_s = \mu_E + \frac{3}{4}(max(E_{obj_i}) - \mu_E) \tag{9}$$

$$thE_i = \mu_E - \frac{4}{5}(\mu_E - min(E_{obj_i})) \tag{10}$$

Again, solidity of the object used to help to the eccentricity define the discrimination.

For this second evaluation, the result is divided into 3 new images, one containing individual chromosomes, other cluster chromosomes and other chromosomes suspects.

A final evaluation is assigned the last image chromosome suspects, for these objects their skeletons are calculated and counted the endpoints.

Endpoints of skeleton are those pixels where each ends the branches of the skeleton of the object. If the skeleton has more than two endpoints indicates that have 2 or more chromosomes in the object it is a group.

The skeleton is the result of thinning the object to obtain a line or sequence of pixels to retain the original shape of the object[4,5]. This is achieved through morphological operators of erosion (ε) and aperture (γ). Then, the skeleton of an object O is defined as (see equations 11 - 14):

$$S(O) = \bigcup_{k=1}^{K} S_k(O) \tag{11}$$

with

$$S_k(O) = \epsilon_B^k(O) - \gamma_B[\epsilon_B^k(O)] \tag{12}$$

where B is the structuring element and $\epsilon_B^k(O)$ indicates k successive erosions of O:

$$\epsilon_B^k(O) = \epsilon_B^k(...(\epsilon_B^2(\epsilon_B^1(O)))) \tag{13}$$

and K is the last iteration before O is the empty set:

$$K = max\{k \mid \epsilon_B(O) \neq \emptyset\} \tag{14}$$

Here there is the case where chromosomes with a centromere subtelocentric have a little separation between their chromatids therefore they could be confused with a cluster. In order to avoid such cases, endpoints are evaluated as follows: if the object has 2 endpoints is claimed to be single, if the object has more than 4 points is said to be a cluster, in other cases, it performs a trimming to the skeleton, which means that removes the branches having a certain number of pixels of the endpoint at the point of intersection of its main branch. To finally have the new endpoints and assign the chromosomes with 2 points to single and the rest to clusters.

The figure 4 depicted the values of the geometric characteristics of some chromosomes.

Type	Object	Solidity	Eccentricity	Area	Endpoints
Cluster		0.5727	0.6657	1091	4
Cluster		0.6150	0.8079	872	3
Single		0.7219	0.3366	553	2
Single		0.8853	0.3340	386	2
Single		0.9136	0.2597	370	2

Fig. 4. Geometric characteristics

4 Result

The dataset used in this work is obtained from http://bioimlab.dei.unipd.it. It is composed of 162 images with 768 × 576 pixels resolution. The images do not necessarily contain a whole set of 46 chromosomes because the image may be spread over different images[2].

Image Q-band are obtained by staining the chromosomes with quinacrine, a fluorescent dye that concentrates in different regions of the chromosomes, giving rise to the characteristic banding patterns that identify the different chromosome types. The images, thus, appear as a dark background onto which the chromosomes stand out with a bright and dark banding.

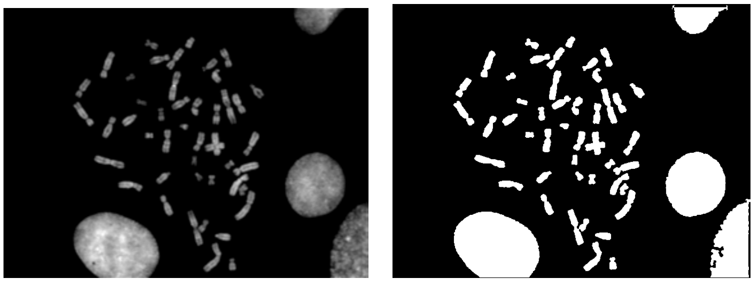
4.1 Extraction of objects of interest

Figure 5 shows the result of applying the "Re-threshold" and perform the extraction of objects of interest on an image containing nuclei.

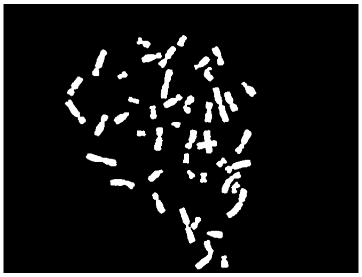
It has two evaluations: 1) evaluates the successful extraction of total images and 2) evaluates the area error of each image.

Evaluation 1: on the the dataset, the method has successfully extracted chromosomes from 148 images , in 3 images was not possible to completely eliminate the unwanted objects and in 11 images some chromosomes were lost (see figure 6).

Evaluation 2: For each image, the are error was obtained with respect to a reference image. This is possible if the result of the proposed method (M) is



(a) Original image (b) Result Rethreshold



(c) Chromosomes extracted

Fig. 5. First step os segmentation

compared with a reference (R) which includes all pixels of interest. View all 4 area error measurements in table1.

Measure	pixel value
True Positive (TP)	$M_{ij} = 1 \wedge R_{ij} = 1$
True Negative (TN)	$M_{ij} = 0 \wedge R_{ij} = 0$
False Positive (FP)	$M_{ij} = 1 \wedge R_{ij} = 0$
False Negative (FN)	$M_{ij} = 0 \wedge R_{ij} = 1$

Table 1. area error measurements

The percentage of area successfully extracted by the method is calculated using the above measures in the following indicators [7]:

$$precision = \frac{TP}{TP + FP} \tag{15}$$

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

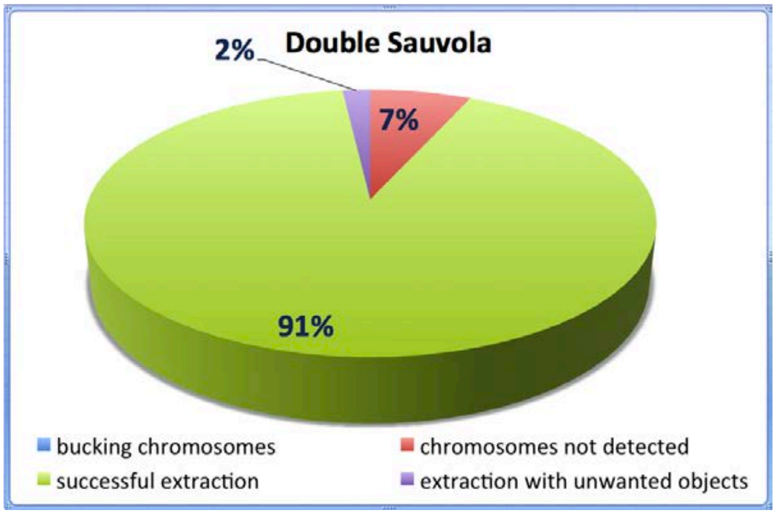


Fig. 6. Result of success of the first segmentation stage of chromosomes.

$$recall = \frac{TP}{TP + FN} \tag{17}$$

$$Fscore = \frac{2 \times precision \times recall}{precision + recall} \tag{18}$$

The dataset has no references images to evaluate the error area. So we take randomly 34 images which are manually binarized to create reference images. The results obtained by the proposed method and adaptive Otsu are evaluated with the reference images. The results are shown in table 2.

Method	**precision**	**recall**	**accuracy**	**Fscore**
Proposed	0.9611	0.8849	0.9924	0.9200
adaptative Otsu	0.9793	0.7778	0.9887	0.8659

Table 2. Evaluation on 20% of the total dataset

The table 2 shows the comparison of the proposed method with adaptive Otsu. The recall indicates that our proposal obtains a greater number of pixels of interest that adaptive Otsu, therefore, the Fscore indicates that the proposed approach has a higher success rate.

4.2 Discrimination

For this step, the resultant binary images of the first stage were used in the discrimination process (93% of images of Q-banding).

For the 151 images that pass this second stage of segmentation has been obtained a script which contains the number of chromosomes in each cluster having one of the images according to the results yielded by the first stage of segmentation. Each image is evaluated and compared with this script and the results obtained are shown in table 3.

Total of clusters	508
Clusters detected	473
Clusters undetected	35
percentage of success	93.5%

Table 3. Result of chromosomes clusters found in the stage of discrimination in Q-banding images.

Groups of chromosomes that correspond to 6.5% were not detected, those who are connected at one end forming a single chromosome, some examples can be seen in figure 7.

Fig. 7. Examples of clusters of chromosomes undetected.

5 Conclusions and future work

According to Poletti et al.[7], adaptive techniques provides the best results in the segmentation as confirmed in this work by applying the proposed technique based on Sauvola thresholding. Obtained results are competitive with others reported in the literature.

The proposed technique of "Double Sauvola" permits to find the appropriate threshold for each pixel, also it avoids cutting the chromosomes in the centromere or any of their bands and it reduces the number of chromosomes together.

The former evidence indicating that the segmentation from the perspective of a local adaptive method gives good results in segmenting images of chromosomes, has been confirmed by obtaining the 93% of successful images of 162 image.

The discrimination method allows to identify unions and overlaps with a 93.5% of success. Although the results are already competitive, we continue working on parameters tuning to rescue the chromosomes corresponding to the FP (chromosomes with separated chromatids) and FN (chromosomes connected at one end) in discrimanción of chromosomes.

As future work the separation of cluster of chromosomes will be explored.

References

1. Gady Agam and Its'hak Dinstein. Geometric separation of partially overlapping nonrigid objects applied to automatic chromosome classification. *IEEE Transactions on pattern analysis and machine intelligence*, 19(11), november 1997.
2. Enrico Grisan, Enea Poletti, and Alfredo Ruggeri. Automatic segmentation and disentangling of chromosomes in q-band prometaphase images. *IEEE Transactions on Information Technology in Biomedicine*, 13(4), july 2009.
3. Enamul Huque. Shape analysis and measurement for the hela cell classification of cultured cells in high throughput screening. Master's thesis, School of Humanities & Informatics University of Skvde, Sweden, February 2006.
4. Sahar Jahani, S. Kamaledin Setarehdan, and Emadedin Fatemizadeh. Automatic identification of overlapping/touching chromosomes in microscopic images using morphological operators. *Machine Vision and Image Processing (MVIP)*, 2011.
5. Ramesh Jain, Rangachar Kasturi, and Brian G. Schunck. *Machine vision*. McGraw-Hill, 1995.
6. L. Ji. Fully automatic chromosome segmentation. *Cytometry*, 17:196–208, 1994.
7. Enea Poletti, Francesca Zappelli, Alfredo Ruggeri, and Enrico Grisan. A review of thresholding strategies applied to human chromosome segmentation. *Computer Methods and Programs in Biomedicine*, 2012.
8. Gunter Rittery and Le Gao. Automatic segmentation of metaphase cells based on global context and variant analysis. *Pattern Recognition*, 41:38–55, 2008.