Automatic segmentation and disentangling of chromosomes in Q-band prometaphase images

Enrico Grisan, *Member, IEEE,* Enea Poletti, and Alfredo Ruggeri*, *Senior Member, IEEE*

***Abstract*—Karyotype analysis is a widespread procedure in cytogenetics to assess the possible presence of genetics defects. The procedure is lengthy and repetitive, so that an automatic analysis would greatly help the cytogeneticist routine work. Still, automatic segmentation and full disentangling of chromosomes are open issues. We propose an automatic procedure to obtain the separated chromosomes, which are then ready for a subsequent classification step. The segmentation is carried out by means of a space variant thresholding scheme, which proved to be successful even in presence of hyper- or hypo-fluorescent regions in the image. Then the tree of choices to resolve touching and overlapping chromosomes is recursively explored, choosing the best combination of cuts and overlaps based on geometric evidence and image information. We show the effectiveness of the proposed method on routine data acquired with different microscope-camera setup at different laboratories: from 162 images of 117 cells totaling 6683 chromosomes, 94% of the chromosomes were correctly segmented, solving 90% of the overlaps and 90% of the touchings. In order to provide the scientific community with a public dataset, the data used in this paper are available for public download.**

***Index Terms*—Chromosome analysis, overlapping chromosomes, adjacent chromosomes, karyotyping, image segmentation.**

I. Introduction

Chromosome karyotyping analysis [1] is an important screening and diagnostic procedure routinely performed in clinical cytogenetic labs. Chromosome are first stained with a fluorescent dye, and then imaged through a microscope for subsequent analysis and classification. Each chromosome in the image has to be identified and assigned to one of 24 classes. The result of this procedure is the so-called karyotype image, in which all chromosomes in a cell are graphically arranged according to an international system for cytogenetic nomenclature (ISCN) [2] classification. Fig. 1 shows a typical PAL resolution (768 x 576, 8 bits/pixel) Q-banding prometaphase image and a karyotype of all chromosomes in that cell.

The appearance of chromosomes depends on the stage of the cell division cycle at which they are viewed. For much of the cell cycle (interphase), individual chromosomes can not be distinguished. They only appear as distinct bodies towards the end of the cycle, at prophase, when they are long string-like objects, contracting and separating at metaphase, just before cell division. Prometaphase is the intermediate stage of contraction between prophase and metaphase. These different stages in the division cycle are characterized by the number of bands visible in the cell, and by the elongation of the chromosomes: the more elongated stages, with a higher resolution of the chromosome structure, are also those that present the greatest difficulty in chromosome analysis, due to the fact that longer chromosomes touch and overlap each other much more frequently than shorter ones.

After automatic chromosome analysis was firstly proposed [3], many years of effort have resulted in the development of commercial cytogenetics systems for analysis of banded chromosome preparations [4]. Most of the studies have concentrated on metaphase chromosomes, avoiding the segmentation difficulties arising from touches and overlaps in the prophase and prometaphase cells. Graham and Piper [1] provide a review of methods used in automated chromosome analysis.

The first step that has to be taken in analysing a chromosome image is the segmentation of chromosomes and chromosome clusters from the image background. The main methods used to segment cytogenetic images are based on the evaluation of a global threshold by means of the Otsu method [5], [6], on a global threshold with a re-thresholding scheme [7], [8], on k-means clustering [9], or finally on the watershed transform [10]. However, the high variability in chromosome fluorescence intensity makes the utilization of a global threshold and of clustering approaches impractical since smaller chromosomes and their terminal parts appear with a lighter intensity than larger ones. Watershed is heavily affected by noise and by the presence of contrast variability. This situation is even worsened by vignetting artifacts, or non-even illumination of the field of view by the UV lamp.

Due to the non-sharp margins of the chromosomes, to the presence of staining debris, or to the fact that long chromosomes touch and overlap, the first segmentation step is usually unable to identify all chromosome as a single object, but rather presents a number of clusters. Some attempts have been made to deal with clusters of touching (but not overlapping) chromosomes [5], [9], [11], [12], and for clusters of overlapping (but not touching) chromosomes [13], [14], where combinations of geometric and densitometric evidence have been used to resolve segmentation ambiguities. Automatic separation of overlapping and touching chromosomes is important for the analysis of prophase and prometaphase images, but has received relatively little attention compared to other aspects of the chromosome analysis problem, such as classification. Ji [7] has proposed methods for automatically segmenting both touching and overlapping clusters that force

The authors are with the Department of Information Engineering, University of Padova, Via Gradenigo 6/a, 35131 Padova, Italy (e-mail: alfredo.ruggeri@unipd.it)

Preliminary results were presented at the 29^{th} IEEE EMBS International Congress, Lyon (F), August, 2007.

This work was partially supported by TesiImaging, Italy.

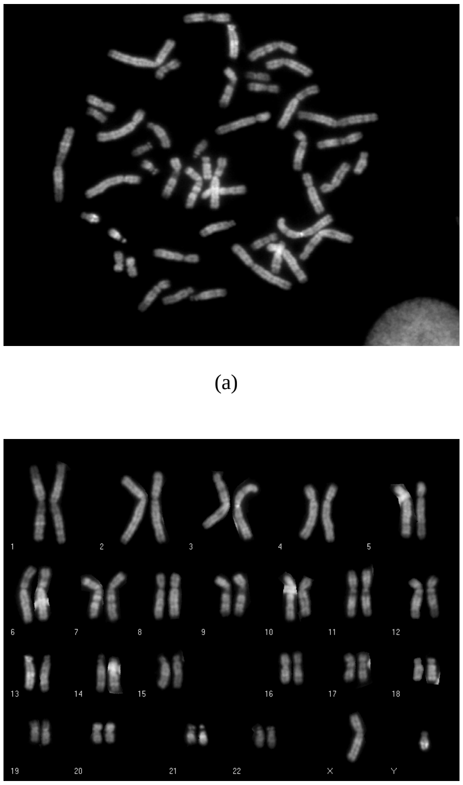
the image to contain a reasonable number of chromosomes (45-47), regardless of their likelihood and of the actual number of chromosomes in the image; the splittig phase is based on *pale paths* [7]. Agam and Dinstein [13] have applied similar reasoning about boundary curvature to separate touching or slightly overlapping chromosomes. Charters and Graham [15] tried to integrate the segmentation and classification steps to resolve overlaps and adjacencies, but trained and tested their algorithm through simulated overlaps only.

Lerner [11] proposed a method to combine the choice of correct cluster disentanglement with the classification stage, resulting in a *classification driven* segmentation. This was however used to separate clusters of two touching chromosomes only. In the same stream, trying to combine the segmentation and classification stage, Ritter [16] proposed a method that relies on the generation of a number of possible chromosomes configuration called *variants*, resulting in a generation of competing sets of karyotypes. The drawbacks of this method are essentially the assumption that there is a limited and known set of configurations onto which the chromosomes may be arranged, the use of an heuristic set of rules to tell a cluster from a single chromosome, and the assumption that an image contains the whole set of cell chromosomes. Moreover, as also mentioned in the paper, given the high number of variants taken into account, it exhibits great difficulty in disentangling clusters with a high number of chromosomes involved, or in dealing with images presenting a low number of single chromosomes identified by the segmentation stage. Finally, the high number of parameters to be set and the long computational time required make its usability in a clinical setting questionable.

We propose here a fully automatic and effective method to segment and disentangle banded chromosomes from Q-band prometaphase images, without any assumption on their number or dimension distribution. To achieve this, we propose a space variant thresholding scheme to segment chromosomes objects from the background. Then we describe a novel yet simple measure of single-chromosome likelihood, to evaluate whether an object is a single chromosome or a cluster, and to score each object. For each cluster, several disentangling hypotheses are generated, evaluated and recursively analyzed, therefore generating an hypothesis tree. The tree of disentangling hypotheses is explored on-line during its generation, with an additional branch-and-bound strategy to keep the computational complexity low.

II. Chromosome data

Q-band images are cytogenetic data 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 bright and dark banding, as shown in Fig. 1a.


(a)

(b)

Fig. 1: Typical Q-band prometaphase image acquired with PAL resolution (a), and the manual kariotyping of the chromosomes (b)

The data set used in this work is composed of 162 images with PAL resolution (576x768 pixels, 8 bits per pixel), of 117 cells acquired during routine laboratory analysis, containing a total number of 6683 chromosomes. The images do not necessarily contain a whole set of 46 chromosomes, as it may happen in routine laboratory acquisition, where the set may be spread over different images. The whole data set is publicly available for download at *http://bioimlab.dei.unipd.it*.

III. Methods

The automatic chromosome identification method we propose is based on a preprocessing stage described in Sec. III-A, whose aim is to separate from the background the chromosomes and chromosome clusters, which are then passed on to the *cluster disentangling* stage. Each object (either chromosome or cluster) is analyzed to test whether it is composed by a single chromosome or if it has to be split by means of a *single-chromosome likelihood*, introduced in Sec. III-D: it is based on the availability of the chromosome axis, whose extraction is described in Sec.III-B, and on a novel measure outlined in Sec. III-C. Each object that does not pass the test becomes the root of a tree, where each node represents one of several cuts and overlaps hypotheses, generated based on geometrical and image evidence, as described in Sec. III-F. Each node is recursively analysed until each leaf of the tree is evaluated as a single chromosome. Then, the best overall combination of hypotheses gives the final disentanglement.

A small set of parameters is used in the algorithm. All parameters have been empirically tuned on 10 randomly chosen images of the available data set.

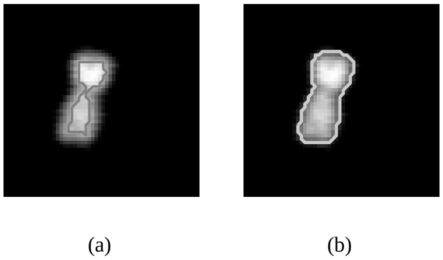


Fig. 2: Particular of the same image segmented using a global threshold obtained with the Otsu method (a), and with the proposed space-variant method (b).

A. Image Segmentation and Cluster Identification

To overcome the drawbacks of some of the methods proposed in the literature and mentioned in Sec. I, we chose to make use of a local adaptive threshold. Hence, we divided the image I in a tessellation of squares of fixed dimension l_{tess} (we used l_{tess}=100 pixels). For each square separately, we evaluated the Otsu threshold [17], which is supposed to best separate the background image component from the foreground. We therefore obtained a matrix of thresholds M_{th} that is scaled by a factor $1/l_{tess}$ with respect to the size of I. Resizing M_{th} to the size of I using linear interpolation, we obtained an image representing the local threshold for each pixel in I. The different results provided by local and global threshold can be appreciated in Fig. 2, where the local threshold is shown to allow a more accurate outline of a sample chromosome.

After identifying and removing by morphological opening small spurious segmented blobs and possible nuclei present in the image, all connected components (blobs) in the segmented image represent either a chromosome or a chromosome cluster. When a blob is considered for analysis, the coordinates (x, y) of all points along its contour are obtained. We can describe the contour as a curve s in a two dimension space, by parametrizing the coordinates by means of a curvilinear coordinate l, and by using a cubic smoothing spline to obtain an interpolated representation with a C^2 continuous function,

$$\boldsymbol{s}(l) = [x(l), y(l)] : [0, 1] \subset \mathbb{R} \rightarrow \mathbb{R}^2 \tag{1}$$

We can then compute the curvature of the contour as:

$$\kappa(l) = \frac{\dot{x}(l)\ddot{y}(l) - \ddot{x}(l)\dot{y}(l)}{(\dot{x}^2(l) + \dot{y}^2(l))^{\frac{3}{2}}} \tag{2}$$

B. Main chromosome axis identification

The identification of the main axis of a blob is of paramount importance, since it is a geometric reference that can be used to extract statistics of mass distribution around the axis, so as to gather information to assess whether the blob is a single chromosome or not. Moreover, all algorithms trying to make use of the banding information for disentangling need to find the main direction of each chromosome segment.

In most works on the subject [11], [18], [19], the *medial axis transform* is used to obtain the main axis of a blob. Unfortunately, on real images and automatically segmented chromosomes, the skeleton obtained through this method presents many spurious branches, which are often difficult to distinguish from the correct ones. Moreover, the criterion used by this skeletonization algorithm always provides a bifurcating axis at both ends of a chromosome.

In order to overcome these limitations, we adopted a modified version of a vessel-tracking algorithm previously proposed for retinal images [20], which is also able to identify bifurcations and crossings. The single axis is obtained by estimating the longest path obtained by the single branches.

C. Accumulation points analysis

Chromosomes may appear in a variety of shapes and configurations, with more or less definitely indented centromeres, making it hard to model the appearance of an average chromosome. However, regardless of the bending and of the centromere indentation, a single chromosome is characterized by smooth sides running along the axis and two caps at both ends of the chromosome shape. Presence of bulges along the sides may very well point out the unlikeliness of the object to be a single chromosome, since bulges suggest the presence of a cap belonging to another chromosome.

We developed a method to detect the presence of convex bulges in an object, assuming that long chromosomes should have two of them corresponding to their extrema, whereas small ones, with a roundish appearance, only one. Starting from the smoothed contour of an object, as described in Sec. III-A, we compute the normal vector $\boldsymbol{n}(l)$ pointing toward the inside of the object. Given a point on the contour with coordinate l_i, we draw a line $\boldsymbol{t}(l_i)$ oriented as $\boldsymbol{n}(l_i)$ of length equal to the median of the chromosome diameters. A weight equal to the curvature $k(l_i)$ is added to every pixel belonging to the line.

The value of each pixel in the resulting image is therefore the integral of the value of these lines weighted by $k(l)$, for all the contour points, so that the peaks visible in the resulting integral image correspond to the centers of rotation of the positively curved portions of the contour (bulges), as can be seen in Fig. 3b and Fig. 3d. Since the lines are weighted by curvature, the different centers of rotation will result in accumulation points with comparable intensities, therefore easily recognizable even when the bulges have different curvature radii.

D. Single-Chromosome Likelihood

The real cornerstone of chromosome cluster disentangling is the criterion used to assess whether a blob can be classified as a single chromosome or as a cluster.

Two main strategies have been proposed to distinguish these situations. The first one is *model-based* [13], where a chromosome silhouette is represented by a properly shaped polygon that has to be fit to the blob under analysis: unfortunately, a model for a single chromosome often produces unreliable

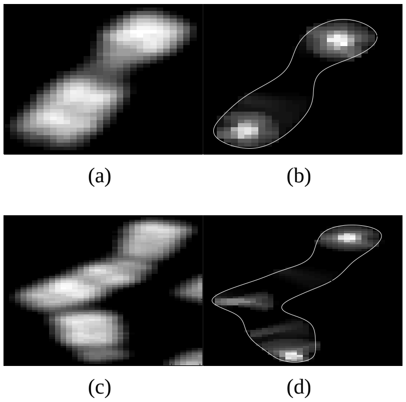


Fig. 3: A single chromosome shows peaks in the accumulator image corresponding to the two chromosome extrema. Panels (a) and (c) show a chromosome and a cluster respectively, panels (b) and (d) show the corresponding accumulator images.

results when fit to a cluster. The second approach proposed is *rule-based* [7], but although simple, this method relies on a set of heuristics and fixed rules to decide whether the blob being analyzed is a single chromosome or a cluster. Moreover, it requires the identification of the blob main sides and diameter, which might not be easy, and it may provide with non-unique solutions when facing chromosome clusters.

We propose here a new *single-chromosome likelihood* (SCL) measure, based on the analysis of the axis curvature and on the presence of sizable bulges detected with the algorithm described in Sec. III-C. Having identified the object axis, we consider that a chromosome can be bent, but very rarely shows a change in curvature sign, resulting in an *S*-like shape. Hence, we mark an object as a cluster if the axis has a change in curvature sign and the difference between the maximum and the minimum curvature is greater than an empirical threshold $th_k = 0.2$.

In the analysis of the accumulator image obtained with the method described in Sec. III-C, we assume that a single chromosome can show two peaks corresponding to bulges in the contour, typically the tips of a chromosome. We then identify the two most prominent peaks and estimate their contrast with respect to the surrounding area. Whenever there are other peaks showing similar contrast, we consider the blob as a cluster, as can be seen in Fig. 3. Hence, given a blob C, with area A_C and with the description of its axis given by the curve $a(l) = [x(l), y(l)]$ onto which the curvature $k_a(l)$ is computed, the score of C is:

$$SCL(C) = \begin{cases} 0 & \text{if bulges or S-shape} \\ 9 - \int k_a(l)dl & \text{if } A_C < th_A \\ 10 - \int k_a(l)dl & \text{otherwise} \end{cases} \qquad (3)$$

so that a blob presenting more than two bulges estimated according to Sec. III-C has score zero, whereas all other blobs have scores depending on the integral of the curvature of their axis. The second condition states that blobs with an area smaller than th_A have a lower score, in order to avoid the splitting of blobs into too many tiny parts. The area threshold th_A was heuristically set at 150 pixels.

E. Generation of candidate splits

As a preliminary analysis for each cluster, by looking at the most prominent local minima of the curvature κ (those with value lower than a threshold $th_\kappa = -0.15$), we derive a set $K = \{k_i, i = 1 : N_k\}$ of points on the contour suggesting the possible presence of touching and overlaps, as previously proposed [12], [13].

Since in complex chromosome clusters, and especially along adjacencies, holes often occur, and holes cannot be present inside single chromosomes, their presence can be exploited as a strong clue that some cuts must be made from the interior of a hole to the outer cluster border. Thus, from each local minimum in K, a path of dark pixels is grown, by iteratively looking at the darkest pixel among the 8-neighbours of the end-pixel of the path. Among all paths ending in a hole, the one with the lower mean intensity of the pixels belonging to the path is chosen, and the cluster is cut along the selected path, so that a cluster without holes results.

1) Touching chromosomes through dark paths: Often, the contact area along adjacent chromosomes has a low but non-zero intensity, due to some fluorescence diffusion from the surrounding chromosomes. Therefore, a path of pixels darker than those belonging to the adjacent chromosomes may be found.

Starting from a point in K, a dark path is grown in the same way as described in in the previous paragraph, with the constraint of ending not in a hole, but on another point in K. The path with the lowest mean intensity is then selected to perform the cut. By this mean, the geometric hints provided by the curvature of the contour function $s(l)$, are combined with the evidence gathered from the image intensity, at variance with [5], [7], [10], where only dark paths (pale in their images) were searched, and with [13], where only geometric evidence was analyzed. This is done in a much simpler way than proposed in [15], since the proposed procedure does not need training, prototypes and hypothesis testing on banding patterns.

2) Overlapping chromosomes: From the set K of selected local minima of the curvature κ of a blob C, all segments entirely contained in C and connecting every two points in K are considered. Since overlaps areas have a rhomboid shape, we look for quadruplets of lines forming a polygon of dimension coherent with the chromosomes diameters and with two pairs of almost parallel sides. Each segment $\boldsymbol{t_{ij}} = [\boldsymbol{k_j} - \boldsymbol{k_i}]$ is defined by its starting and ending points k_i and k_j, and its orientation ϑ_{ij} can be easily derived. The distance between two segments $\boldsymbol{t}_{ij}$ and $\boldsymbol{t}_{mn}$, can be defined as in [13] using the metrics:

$$d(\boldsymbol{t}_{ij}, \boldsymbol{t}_{mn}) = \frac{\boldsymbol{t}_{ij} \times \boldsymbol{t}_{jm}}{|\boldsymbol{t}_{ij}|} \qquad (4)$$

All pairs of segments whose distance is less than $th_C = 1.2 * ad$ pixels, and whose orientations differ by less than $\pi/6$ are considered; ad is the average diameter of the single chromosomes in the image. If no single chromosome has been found yet, ad is set to the empirical value of 45 pixels.

Then, with a greedy approach, all lines completely included in C and forming a polygon close to a rhomboid are selected, the region they delimit is considered to be an overlap, and the cuts to obtain the separated chromosomes are performed along the lines $\{\boldsymbol{t}_{jn}, \boldsymbol{t}_{mi}\}$ and $\{\boldsymbol{t}_{ij}, \boldsymbol{t}_{mn}\}$.

3) Touching chromosomes through geometric cuts: Finally, cut lines suggested by geometrical cues are searched. Candidate cut lines are those lines linking two points in K that are entirely inside the blob. Since their number may be large, this last procedure may be computationally heavy. In principle, for an exhaustive search on N candidate cut lines, the evaluation of $\sum_{i=1}^{N-1}\binom{N}{i}$ combinations of lines is required, and for every combination the computation of SCL for each blob resulting after the application of the selected cuts is also required.

Since we assume that most of the possible disentanglements may have been performed through the previous methods, we expect very few possible cuts to be left unexplored. Hence, we restrict the maximum number of chromosomes in the remaining clusters to be 3 at most, looking then for a maximum combination of 2 simultaneous cuts.

The combinations that result with the minimum number of lines and the maximum sum of SCL over the obtained blob will be selected to generate the splitting hypothesis for the cluster.

F. Generation and exploration of the hypotheses tree

When a cluster has to be cut into its constituent chromosomes, a number of reasonable splits (overlaps, paths or cuts) are usually available. If no assumption on the minimal dimension of a chromosome or on the maximum number of operations for separating the cluster has been made, a set of small very regularly shaped chunks would be obtained. To avoid this situation, we generate competing combinations of hypotheses of how the cluster is made up from single chromosomes, and we evaluate each combination using an Akaike-like model selection criterion [21]. We therefore choose as the more likely hypothesis the one that provides the best combination of high SCL scores for all identified chromosomes, together with the most parsimonious combination, resulting in the disentanglement of a cluster with the minimum number of constituent chromosomes.

Given a set of N chromosomes $C_M = \{c_i; i = 1 \dots N\}$ that results from the disentanglement of a cluster with M splits, its likelihood is evaluated as:

$$\Lambda(C_M) = w_1 \cdot log\left(\frac{1}{N}\sum_{i=1}^{N} SCL(c_i)\right) + w_2 \cdot logM \qquad (5)$$

where w_1 and w_2 are weights summing to 1 and empirically set to 0.2 and 0.8 respectively in our implementation. Starting from an initial cluster B, the possible cuts and overlaps are identified as described in Sec. III-E1, Sec. III-E2 and Sec. III-E3, thus generating different hypotheses along the disentanglement path. Each generated hypothesis splits the cluster into two blobs, that will be recursively evaluated until all blobs achieve a SCL score greater than 0, indicating that they are indeed single chromosomes. We eventually obtain a pool of L sets $C^l_{M_l} = \{c_{i,l}; i = 1 \dots N_l\}$, with $l \in [1, \dots L]$, of complete hypotheses on how to disentangle the cluster. The one obtaining the maximum value of Λ is retained as the more likely:

$$C_M = \arg\max_l \Lambda(C^l_{M_l}) \qquad (6)$$

TABLE I: Clusters presenting overlapping chromosomes correctly resolved, compared to results reported in the literature. It has to be noted that the clusters analyzed in [22] are simulated from manually segmented single chromosomes, hence their appearance might be different from natural clusters.

Method	Overlaps	Resolved
Agam et al. (1997) [13]	124	82%
Charters et al. (1999) [22]	136	71%
Garcia et al. (2003) [14]	200	62%
Proposed	201	90%

An example of a complete hypothesis tree generated for a cluster, is shown in Fig. 4.

The recursive application of Sec. III-E1,Sec. III-E2 and Sec. III-E3 to a cluster, with the stemming of branches every time a new cluster is obtained as result of a split, grows an hypothesis tree in which the leafs are the identified candidate chromosomes, and the complete sets $C^l_{M_l} = \{c_{i,l}; i = 1 \dots N_l\}$ of constituent chromosomes are obtained by exploring the tree and including into the same set all leafs that are linked to a visited node.

The building of the entire hypothesis tree can become computationally cumbersome, as clusters with many chromosomes may be analyzed. Moreover, during the recursive generation of the hypothesis tree, there are subtrees that can not compete with the best solution already found according to Eq. 5, no matter what the scores of their leafs are. This happens when the number of splits outweights the goodness of the chromosomes SCL. Hence, to reduce the computational burden in the generation of the tree, we use a branch-and-bound strategy that keeps track for each node of the best solution for the whole disentanglement already found, with a score Λ_{max}, and the number of splits required to reach that node. We evaluate the combination split score Λ of the path leading to that node, assuming that the remaining cluster will be divided into two perfect chromosomes, i.e. into two blobs with maximum SCL score. If the value of Λ is greater than Λ_{max}, the node is considered and the subtree is generated and explored, otherwise the subtree will be pruned, and the path leading to the node discarded.

IV. Results and Discussion

By using the algorithm described, we analyzed 162 images containing a total of 6683 chromosomes. All images had also their chromosomes manually identified by an expert cytologist. Along with the visual appearance of the results, an example of whom is shown in Fig. 5, we are interested in evaluating the ability of the proposed algorithm to correctly separate clusters

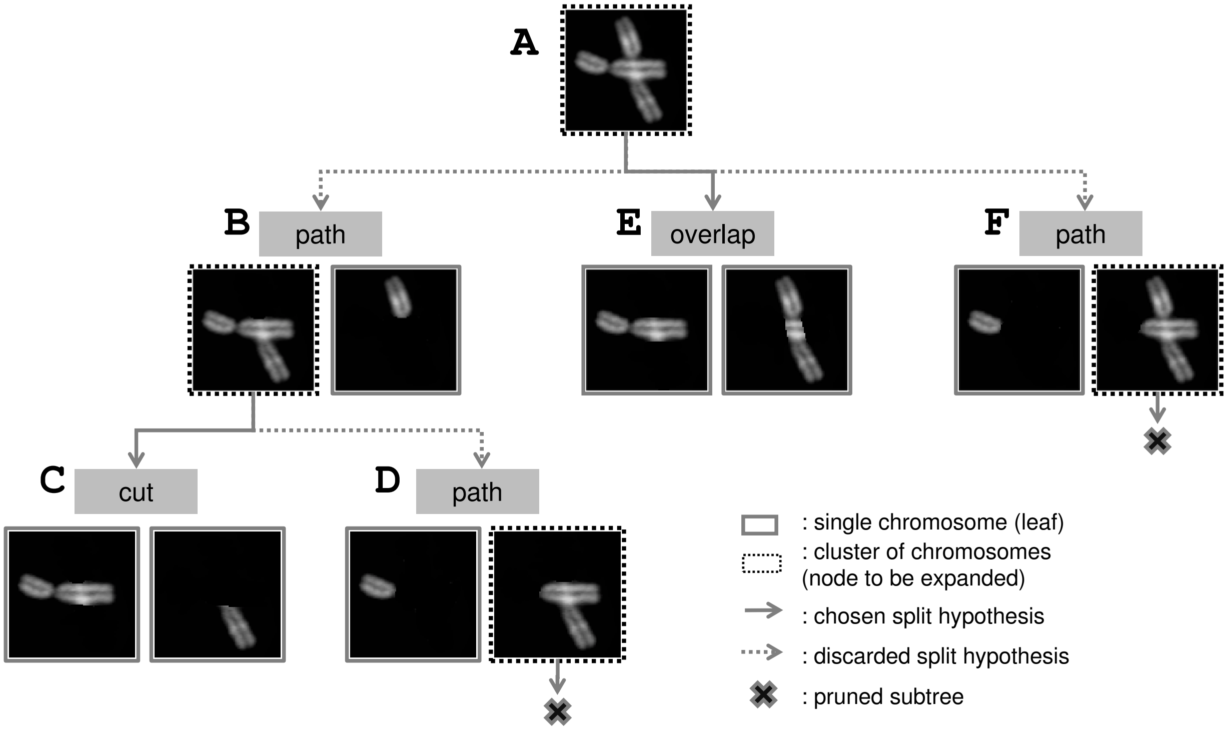

A
B
path
E
overlap
F
path
C
cut
D
path
: single chromosome (leaf)
: cluster of chromosomes (node to be expanded)
: chosen split hypothesis
: discarded split hypothesis
: pruned subtree

Fig. 4: The disentanglement hypothesis tree generated for the cluster A. The algorithm traverses this search tree recursively, from the root down, in depth-first order (A, B, C, D, E, and F) checking at each node the SCL of the obtained parts. When all the leaves of a node are found, the fitness Λ related to that node is calculated (Sec. III-F). For the not-single chromosomes, the best obtainable fitness Λ is consistently estimated and compared to the current one to provide the branch-and-bound strategy.

into the composing chromosomes. Different indexes may be obtained to assess the ability to resolve clusters with overlaps and clusters with adjacencies. In Tab. I we report the fraction of overlaps correctly resolved with respect to the manually identified overlappings, whereas in Tab. II we show in the same way the fraction of correct separations of adjacencies with respect to their total number. In the tables we also report for comparison all similar results reported in the literature, to the best of our knowledge.

TABLE II: Clusters presenting adjacent chromosomes correctly resolved, compared to results reported in the literature

Method	Adjacencies	Resolved
Ji (1989) [5] set 1	458	95%
Ji (1989) [5] set 2	565	98%
Lerner (1998) [11]	46	82%
Shunren et al. (2003) [12]	40	92%
Proposed	819	90%

Finally, as the objective of our algorithm is to correctly identify all 6683 chromosomes separately, we report in Tab. III the fraction of correctly identified single chromosomes, along with the results reported in some previous papers. It must be emphasized that the data sets are different for each study, except those from the same author (as for [5] and [7]); but since the number of chromosome involved in the evaluation is quite high, at least a few hundreds, a general idea of the performance of the different systems might be obtained. It is worth noting that most of the algorithms previously proposed are not tested on real images, but rather on simulated overlaps obtained by combining manually segmented chromosomes [15], [22], or by analyzing manually segmented clusters [11]–[13]. Only [5], [7], [10] segment raw images, even if they restrict themselves to separate only touching chromosomes, without solving the whole disentanglement problem.

From these results, the recursive approach we propose, together with the local threshold, appears to be able to provide the most thorough segmentation of chromosomes, with performance similar to the algorithm proposed in [7] with respect to single-chromosomes identification.

Even if in solving adjacencies the proposed algorithm is not the best method overall, it still achieves very good performance, at the same time obtaining far better performance than other methods in resolving overlapping chromosomes, and still maintaining a very reasonable computational complexity.

The critical point of any algorithm trying to segment single chromosomes and resolve clusters, is the metric used to evaluate if a blob is a single chromosome rather than a cluster: the measure we propose is computationally simple and stable, at the same time providing an effective measure of chromosome-like shape for a blob.

Another issue worth noting is that there is no way with the present algorithm to tell an overlap from an adjacency when one of the chromosomes involved does not have a sizable part of itself on both sides of the overlap site. An hyperfluorescent region might be a clue for distinguishing between the two

cases, but this is not always the case.
The average running time of the prototype developed in MatLab® on a PC with a Pentium 4, 3.2GHz, is 6 minutes, with a maximum time of 14 minutes to process a particularly complex image with a single cluster of 24 chromosomes.

TABLE III: Chromosome correctly segmented, compared to results reported in the literature

Method	Chromosomes	Correct
Popescu et al. (1999) [23]	460	82%
Ji et al. (1994) set 1 [7]	11279	95%
Ji et al. (1994) set 2 [7]	19719	91%
Agam et al. (1997) [13]	1150	94%
Karvelis et al. (2005) [10]	940	93%
Proposed	6683	94%

V. Conclusions

We have presented an algorithm able to automatically identify chromosomes in prometaphase images, taking care of a first segmentation step and then of the disentanglement of chromosome clusters by resolving separately adjacencies and overlaps with a greedy approach, which ensures that at each step only the best split of a blob is performed.
The performance of the proposed methods are better or comparable to the best of other methods reported in the literature. To the best of our knowledge, this is the first method to simultaneously tackle segmentation, overlaps and adjacencies, providing a tool able to automatically analyze an image, and whose results can be handed over with minimal human intervention to a classifier for automatic karyotyping.
Additionally, we made the entire image set used in the paper publicly available on the web, to encourage the development and evaluation of competing algorithms onto the same data set.

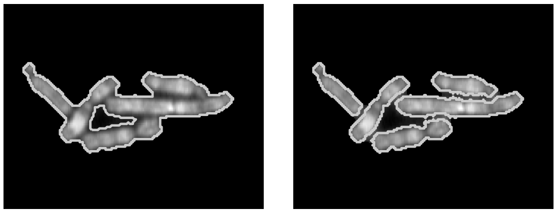

(a) (b)

Fig. 5: Starting from the cluster depicted in (a), the result of the disentanglement for is shown in (b). All chromosomes have been correctly segmented.

Acknowledgment

The authors wish to thank TesiImaging SrL for financial support and for having kindly provided chromosome images and manual karyotypes.

References

[1] J. Graham and J. Piper, *Automatic karyotype analysis*, 1994, pp. 141–185.

[2] Intarnational Standing Committee on Human Cytogenetic Nomenclature, *ISCN: an international system for human cytogenetic nomenclature (2005)*, L. G. Shaffer and N. Tommerup, Eds. Karger and Cytogenetics and Genome Research, 2005.

[3] R. Ledley, "High speed automatic analysis of biomedical pictures," *Science*, vol. 146, pp. 216–223, 1964.

[4] C. Lundsteen and A. Martin, "On the selection of systems for automated cytogenetic analysis," *American Journal of Medical Genetics*, vol. 32, pp. 72–80, 1989.

[5] L.Ji, "Intelligent splitting in the chromosome domain," *Pattern Recognition*, vol. 22, no. 5, pp. 519–532, 1989.

[6] L. V. Guimarães, A. Schuck Jr., and A. Elbern, "Chromosome classification for karyotype composing applying shape representaion on wavelet packet transform," in *The 25th Silver Anniversary International Conference of the IEEE Engineering in Medicine and Biology Society*, Cancun (Mx), September 2003, pp. 941–943.

[7] L. Ji, "Fully automatic chromosome segmentation," *Cytometry*, vol. 17, pp. 196–208, 1994.

[8] R. J. Stanley, J. M. Keller, P. Gader, and C. W. Caldwell, "Data-driven homologue matching for chromosome identification," *IEEE Transactions on Medical Imaging*, vol. 17, no. 3, pp. 451–462, June 1998.

[9] B. Lerner, "Toward a completely automatic neural-network-based human chromosome analysis," *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, vol. 28, no. 4, pp. 544–552, August 1998.

[10] P. S. Karvelis, D. I. Fotiadis, M. Syrrou, and I. Georgiou, "Segmentation of chromosome images based on a recursive watershed transform," in *Proceedings of the 3rd European Medical and Biological Engineering Conference, EMEBC'05*, Prague (CZ), November 2005.

[11] B. Lerner, H. Guterman, and I. Dinstein, "A classification-driven partially occluded object segmentation (CPOOS) method with application to chromosome analysis," *IEEE Transactions on Signal Processing*, vol. 46, no. 10, pp. 2841–2847, October 1998.

[12] X. Shunren, X. Weidong, and S. Yutang, "Two intelligent algorithms applied to automatic chromosome incision," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. (ICASSP '03).*, April 2003, pp. 697–700.

[13] G. Agam and I. Dinstein, "Geometric separation of partially overlapping nonrigid objects applied to automatic chromosome segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 1212–1222, November 1997.

[14] C. Urdiales García, A. Bandera Rubio, F. Arrebola Pérez, and F. Sandoval Hernández, "A curvature-based multiresolution automatic karyotyping system," *Machine Vision and Applications*, vol. 14, pp. 145–156, 2003.

[15] G. C. Charters and J. Graham, "Disentangling chromosome overlaps by combining trainable shape models with classification evidence," *IEEE Transactions on Signal Processing*, vol. 50, no. 8, pp. 2080–2085, August 2002.

[16] G. Ritter and L. Gao, "Automatic segmentation of metaphase cells based on global context and variant analysis," *Pattern Recognition*, vol. 41, pp. 38–55, 2008.

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

[18] M. Moradi and S. K. Staredhan, "New featyres for automatic classification of human chromosomes: a feasibility study," *Pattern Recognition Letters*, vol. 27, pp. 19–28, 2006.

[19] J. Piper and E. Granum, "On fully automatic feature measurement for banded chromosome classification," *Cytometry*, vol. 10, pp. 242–255, 1989.

[20] E. Grisan, A. Pesce, A. Giani, M. Foracchia, and A. Ruggeri., "A new tracking system for the robust extraction of retinal vessel structure," in *Proc. 26th Annual International Conference of IEEE-EMBS*. New York: IEEE, 2004, pp. 1620–1623.

[21] H. Akaike and F. Sandoval Hernández, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[22] G. C. Charters and J. Graham, "Trainable grey-level models for disentangling overlapping chromosomes," *Pattern Recognition*, vol. 32, pp. 1335–1349, 1999.

[23] M. Popescu, P. Gader, J. Keller, C. Klein, J. Stanley, and C. Caldwell, "Automatic karyoptyping of metaphase cells with overlapping chromosomes," *Computers in Biology and Medicine*, vol. 29, pp. 61–82, 1999.