Geometric Correction of Deformed Chromosomes for Automatic Karyotyping

Shadab Khan*, Alisha DSouza*, João Sanches and Rodrigo Ventura

** (Asterisk indicates equal contribution)*

***Abstract*— Automatic Karyotyping is the process of classifying chromosomes from an unordered karyogram, into their respective classes to create an ordered karyogram. Automatic karyotyping algorithms typically perform geometrical correction of deformed chromosomes for feature extraction; these features are used by classifier algorithms for classifying the chromosomes. Karyograms of bone marrow cells are known to have poor image quality. An example of such karyograms is the Lisbon-K_1 (LK_1) dataset, which is used in our work. Thus, to correct the geometrical deformation of chromosomes from LK_1, a robust method to obtain the medial axis of the chromosome was necessary. To address this problem, we developed an algorithm that uses the seed points to make a primary prediction. Subsequently, the algorithm computes the distance of boundary from the predicted point, and the gradients at algorithm-specified points on the boundary to compute two auxiliary predictions. Primary prediction is then corrected using auxiliary predictions, and a final prediction is obtained to be included in the seed region. A medial axis is obtained in this way, which is further used for geometrical correction of the chromosomes. This algorithm was found capable of correcting geometrical deformations in even highly distorted chromosomes with forked ends.**

I. Introduction

Automatic Karyotyping is the process of ordering and classifying the chromosomes into their respective classes; 22 pairs of autosomes and a pair of allosomes. Ordered karyograms are created using karyotyping, which are used to study chromosomal morphology. These studies are useful for detection of diseases, particularly cancer, such as leukemia. By identifying the aberration in chromosome features, such as, position of centromere, length of chromosome, area of the chromosome and band pattern etc., clinicians are able to judge whether the samples contain signatures of a disease. Automatic karyotyping algorithms extract features of chromosomes, and use those features to classify the chromosomes. However, karyotyping of chromosomes from bone marrow cells poses a challenging task due to poor details in images; a feature typical of unordered karyograms of bone marrow cells. The data set that we are working on, is a set of ordered karyograms of bone marrow cells, and is called Lisbon-K_1 (LK_1) [1]. Thus, extraction of features from LK_1 chromosomes is a challenging problem.

Several features of chromosomes are used for the classification of chromosomes. Some of the most prominent features are band profile, position of centromere and dimension of the chromosomes. However, chromosomes from LK_1 are inadequately condensed and elongated for reliable identification of the centromere position. Thus, an accurate band profile of chromosome becomes even more important [2]-[5]. Band profile computation, in turn requires an accurate geometric correction of chromosomal deformations. In previous studies, several algorithms for geometric correction of chromosomes have been presented. Use of MAT [1],[6]-[7] and infinite thinning [8] has been previously used to obtain a medial axis to correct the shape of the chromosome. Different methods of geometric correction using vessel-tracking algorithm [4], and by segmenting the chromosome into polygons have also been proposed [3]. Most of these algorithms obtain an initial guess and extrapolate it to obtain the medial axis, which is then used for geometric correction. The extrapolation techniques overlook the variations in the boundary and rely solely on the seeds, thus introducing inaccuracies in medial axis towards the ends of the chromosome, which in turn affects the geometrical correction.

The motivation of our work was to reduce these inaccuracies and to extract more accurate features for the classification of chromosomes. We previously developed an algorithm that obtains the initial seed region by pruning the skeleton of the chromosomes [9]. The seed region was then extrapolated. To account for the variations in the boundary, the algorithm kept track of the distances of extrapolated point, from the boundary of chromosomes. While this algorithm worked well, it had two shortcomings: 1) In the cases where chromosomes had "forked" towards the end, the medial axis wasn't obtained in such a way that it could capture the forked portions, 2) While the medial axis, and band profile were computed with high accuracy, the algorithm couldn't correct the deformation in chromosome shapes with as much fidelity as is necessary. With our new algorithm, we have addressed these issues. In addition to the primary prediction and distances from the boundary, the algorithm considers the gradients along the boundary to extrapolate the seed region. This leads to improvements in band profiles, and geometrical. This paper, describes the working of our new algorithm.

II. Method

To accomplish geometric correction, the algorithm has three main sections: 1) Seed Region Extraction, 2) Medial Axis Estimation, 3) Axis Smoothing and Geometric Correction. These are described in order:

Shadab Khan is with the Thayer School of Engineering, Dartmouth College, Hanover, NH, USA (e-mail: shadab.khan.th@dartmouth.edu).

Alisha DSouza is with the Thayer School of Engineering, Dartmouth College, Hanover, USA (e-mail: Alisha.V.D'souza.TH@Dartmouth.edu).

João Sanches is with the Department of Bioengineering, Instituto Superior Técnico, Lisbon, Portugal (e-mail: jmrs@ist.utl.pt).

Rodrigo Ventura is with the Department of Electrical and Computer Engineering, Instituto Superior Técnico, Lisbon, Portugal (e-mail: rodrigo.ventura@isr.ist.utl.pt).

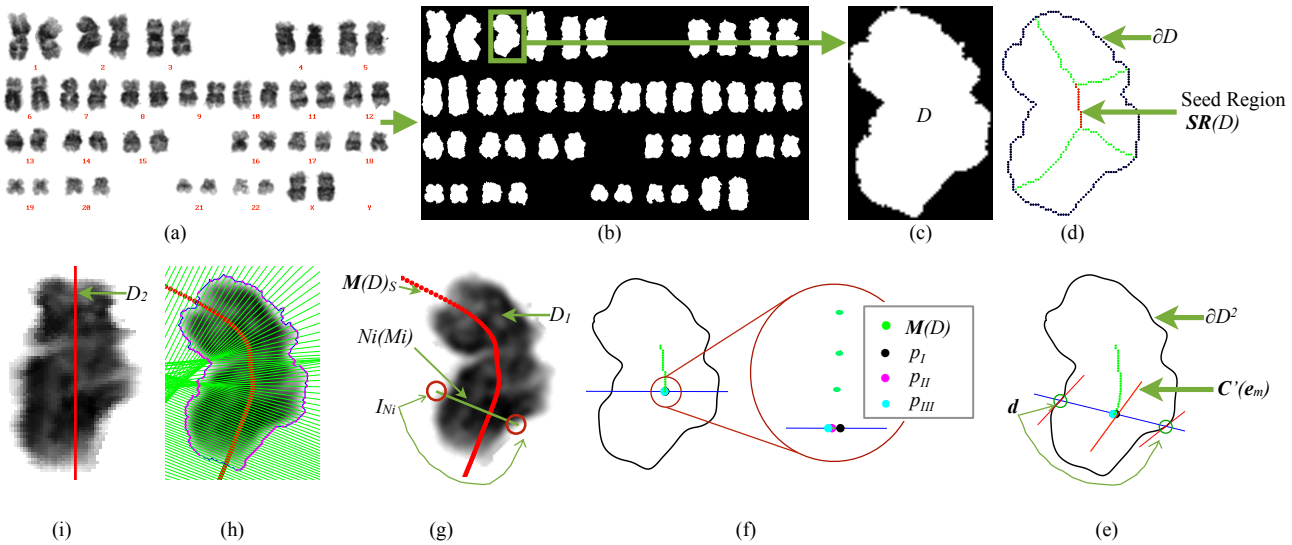


Figure 1. (a) Ordered Karyogram, (b) Binarized and segmented karyogram, bounding box is shown for one of the chromosomes, (c) Extracted chromosome, (d) Chromosome with skeleton, seed region is marked, (e)-(h) Annotated chromosome processing stages, (i) Geometrically corrected chromosome, obtained from (h), correspondence between chromosomes from (g) and (i) can be observed.

1) *Seed Region Extraction*

The karyograms in LK_1 dataset are ordered. The algorithm begins with the extraction of chromosome, which is then processed in subsequent steps. The karyogram is first binarized and segmented. Connected components in the segmented karyogram represent chromosome, and are extracted by calculating the dimensions of a bounding box that encloses each chromosome. Chromosome extraction is followed by skeleton computation, which is further used for seed region extraction.

Since the chromosomes from LK_1 dataset have highly irregular boundary, existing algorithms produce skeletons with too many branches. Thus, an algorithm developed by X. Bai et al. [10] was chosen, which can generate a skeleton with desired number of branches. The algorithm is fast, and robust with respect to irregularities in the boundary of the input shape. To help in the further discussion of the algorithm, few definitions are noted below. To help the reader in relating the notations to chromosome image, Fig. 1 shows several intermediate steps of geometrical correction of chromosomes with annotations.

Let us define a 2-dimensional space $\mathbb{R}^2$ containing a connected subset D that has a boundary ∂D constituting of analytic closed curves, Fig. 1 (c)-(d). Then, the skeleton $\boldsymbol{S}(D)$ of a set D is the locus of the center of a disk that touches ∂D and is independent of other disks in D [11]. $T(s)$, is a set resulting from operation $\{ \partial D \cap \text{Disk}(s) \}$, where Disk($s$) is a maximal disk centered at s where $s \in \boldsymbol{S}(D)$. Degree of s, deg(s) is defined as cardinality of $T(s)$. Then, the bifurcation points of $\boldsymbol{S}(D)$ are defined as $b := \{s \in \boldsymbol{S}(D) : \deg(s) \geq 3\}$. An end point is defined as $e := \{ \boldsymbol{S}(D) \cap \partial D \}$. The algorithm described in [10] returns a skeleton with 4 branches, 4 end points and 2 bifurcation points. Then, Seed Region, $\boldsymbol{SR}(D)$, is defined as $\boldsymbol{SR}(D) := \{s \in \boldsymbol{S}(D) : s$ is between $b\}$ and is obtained from the skeleton of the chromosome, Fig. 1 (d).

2) *Medial Axis Estimation*

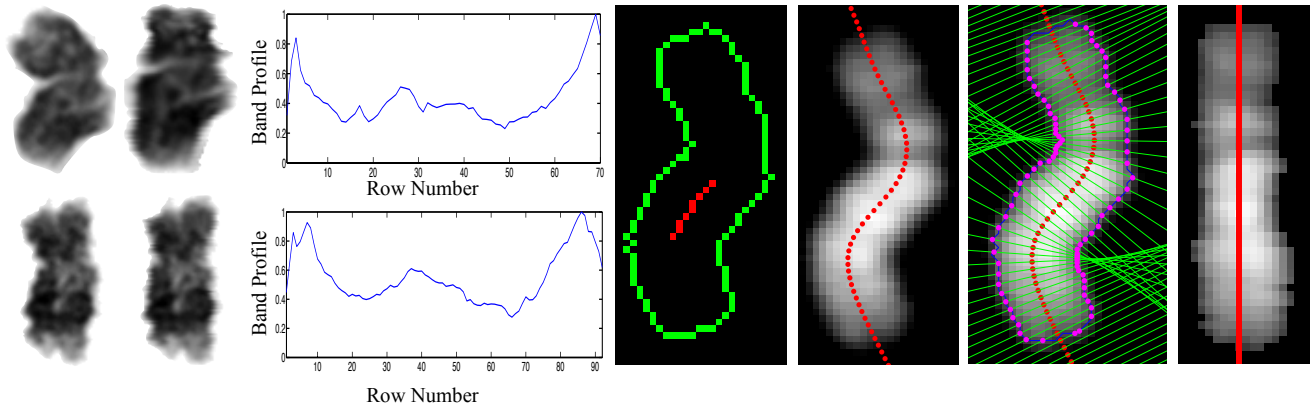
Once the seed region has been obtained, it is extrapolated into a medial axis. To accomplish this, the boundary is smoothened by first fitting a piecewise cubic spline to ∂D and using regression to find the smooth boundary, ∂D^2, Fig. 1 (e). ∂D^2, is then differentiated with respect to $\boldsymbol{x}$, at all $x \in \partial D^2$, to estimate the boundary derivative $\partial D^2{}'$. Medial axis $\boldsymbol{M}(D) \equiv [\boldsymbol{M}_x \boldsymbol{M}_y]$ is then defined as an axis of symmetry obtained by extrapolating the seed region, so that $\boldsymbol{M}(D)$ traverses D and $\boldsymbol{M}_x$ is nonstrictly increasing with respect to $\boldsymbol{x}$, Fig. 1 (g). Note that for a given vector $\boldsymbol{V}$, $\boldsymbol{V}_x$ and $\boldsymbol{V}_y$ refer to its components in the x and y directions respectively. Further, $\boldsymbol{f}'$ is assumed to be the derivative of $\boldsymbol{f}$ with respect to x. Extrapolation from $\boldsymbol{SR}(D)$ to $\boldsymbol{M}(D)$ is performed using the rules described below.

To grow $\boldsymbol{M}(D)$ is to append a new element $\boldsymbol{e}_M$ such that : if $\boldsymbol{C}$ is the curve describing the spatial distribution of $\boldsymbol{M}(D)$, then $\boldsymbol{C}'(\boldsymbol{e}_M)$ is the tangent to $\boldsymbol{C}$ at $\boldsymbol{e}_M$ and $norm(\boldsymbol{C})$ at $\boldsymbol{e}_M$ is the normal to $\boldsymbol{C}'(\boldsymbol{e}_M)$ at $\boldsymbol{e}_M$, Fig. 1(e). Let $\boldsymbol{d} := \{d \in \boldsymbol{C} : d \in \boldsymbol{C} \cap norm(\boldsymbol{C})$ at $\boldsymbol{e}_M\}$ be the set of points that describe the intersection of the normal to $\boldsymbol{C}$ and $\boldsymbol{C}$, Fig.1(e). Subsequently, $\boldsymbol{e}_M$ is a valid point as long as it satisfies all or one of the conditions described below (cannot be generalized to all D):

Condition 1: $\| \boldsymbol{e}_M - \mu_{\boldsymbol{d}} \| \leq \psi$, where ψ is the error limit; here "$\| \cdot \|$" operator is the l_2- norm and $\mu_{\boldsymbol{d}}$ is the midpoint of line connecting the points in $\boldsymbol{d}$.

Condition 2: $\boldsymbol{C}'(\boldsymbol{e}_M) \approx$ mean $(\partial D^2{}'(\boldsymbol{d}))$; where $\partial D^2{}'(\boldsymbol{d})$ is the gradient of the smoothened boundary $\partial D^2{}'$ at the points in $\boldsymbol{d}$. This condition ensures that the gradient of $\boldsymbol{C}$ at $\boldsymbol{e}_M$ varies with the variations in the boundary $\partial D^2{}'$ at the intersection points $\boldsymbol{d}$. This is an intuitive measure and follows from the idea that we need $\boldsymbol{M}(D)$ to be as spatially dynamic as the boundary ∂D^2.

Condition 3: $\| \boldsymbol{e}_M - \boldsymbol{e}_{(1 \text{ or } 2)} \| \leq \gamma$, where $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ are the end points of $\boldsymbol{M}(D)$ before inclusion of $\boldsymbol{e}_M$, and γ is a threshold parameter that ensures that $\boldsymbol{e}_M$ lies in the vicinity of



(a) (b)

Figure 2. (a) Band Profile of chromosomes from the same class, straightened chromosome is shown along with the original deformed chromosome (b) A case of a chromosome with small seed region that was accurately corrected.

the end points of $\boldsymbol{M(D)}$. If this condition is violated, unwanted irregularities may be induced.

The algorithm estimates $\boldsymbol{e}_M$ as a weighted sum of a primary prediction P_I and two auxiliary predictions P_{II} and P_{III}, which are obtained using a 3-step process described below.

Step 1 : To begin the algorithm, assign $\boldsymbol{M(D)} = \{s \in \boldsymbol{SR(D)}\}$. A training set S is formed by sampling N_p ($N_p = 6$) points at the extremities of $\boldsymbol{M(D)}$. The primary prediction, P_I is obtained by the technique described in our previous work [9]. Since $\boldsymbol{M}_x$ is assumed to be nonstrictly increasing with respect to x, let $P_{Ix} = x$, where x is the x-coordinate of the next $\boldsymbol{e}_M$ to be appended to the $\boldsymbol{SR(D)}$ A hypothesis h_θ is then defined as, $h_\theta(x) = \theta_0 + \theta_1 x$, where $h_\theta(x)$ is the function used to predict y-coordinates of P_{Iy} for input P_{Ix}. Hypothesis, $h_\theta(x)$, is calculated by fitting a weighted linear polynomial to $\boldsymbol{S}$ as described in [9]. Once $h_\theta(x)$ is available, P_{Iy} is given by $P_{Iy} = h_\theta(P_{Ix})$. This method of prediction ensures that Condition 1 is satisfied for all cases with low value of ψ.

Next, for estimating the auxiliary predictions P_{II} and P_{III}, points of intersection of the orthogonal to $\boldsymbol{C}$ at P_I (called $\boldsymbol{norm(C)}$ at P_I) and ∂D^2 are required. The points of intersection are $\boldsymbol{d}$.

To calculate P_{II}, the x-coordinate of P_{IIx} (x-coordinate of P_{II}) is set to be $P_{IIx} = x$, and the y-coordinate P_{IIy} is assigned the mean of the y-coordinates of points in $\boldsymbol{d}$ ($\boldsymbol{d} \equiv [\boldsymbol{d}_x\ \boldsymbol{d}_y]$). Then,

$$P_{II} = [\, x\ \ \mu_{dy}\,] \tag{1}$$

where $\mu_{\boldsymbol{dy}}$ is the mean of $\boldsymbol{d}_y$.

To calculate P_{III}, the derivatives of ∂D^2 at the points $\boldsymbol{d}$ are considered. These are represented by $\partial D^2\text{'}(\boldsymbol{d})$. The x-coordinate of P_{III}, P_{IIIx}, is set to be $P_{IIIx} = x$, and its y-coordinate P_{IIIy} is calculated as follows. Further, prediction P_{III} is required to satisfy Condition 3 . This means that a line joining the end point $e_{(1\text{ or }2)}$ of $\boldsymbol{M(D)}$, to P_{III} has a gradient that is a function of the gradients of boundary at the points in $\boldsymbol{d}$. This line, $h_\xi(x)$, is obtained using equations (2)-(4)

$$h_\xi(x) = \xi_0 + \xi_1 x \tag{2}$$

$$\xi_1 = \text{mean}\,(\partial D^2\text{'}(\boldsymbol{d})) \tag{3}$$

$$\xi_0 = e_{iy} - \xi_1 e_{ix}\,;\ \text{for i = 1 or 2} \tag{4}$$

Here ξ_1 is the slope of the line $h_\xi(x)$, and ξ_0 is its y-intercept. P_{IIIy} is assigned the value $h_\xi(x)$ and hence,

$$P_{III} = [\, x\ \ h_\xi(x)\,] \tag{5}$$

We have all three predictions: P_I, P_{II} and P_{III}, Fig. 1 (f).

Step 2: The auxiliary predictions are validated by checking if $\| P_I - P_{II} \| \leq \text{TOL}$ (TOL is set to a default of 1.5). This check ensures that prediction doesn't lie outside the expected region; it's done to suppress unexpected deviations in $\boldsymbol{M(D)}$. Note that the algorithm checks only for P_{II} to be in the vicinity of P_I. If the inequality is true, then P_{II} and P_{III} are valid and the algorithm continues. If the inequality is not true, then: $\boldsymbol{e}_M = P_I$. Once P_{II} and P_{II} have been validated, $\boldsymbol{e}_M$ is estimated as a weighted mean of the 3 predictions:

$$\boldsymbol{e}_M = (W_I \times P_I + W_{II} \times P_{II} + W_{III} \times P_{III}) / (W_I + W_{II} + W_{III}) \tag{6}$$

The weight vector $\boldsymbol{W} = [\, W_I\ W_{II}\ W_{III}\,]$ is assigned a default value of [1 1 1] and can be modified to suit specific cases where the boundary ∂D is too irregular to be used with default weights. Such a weighting allows more control over the seed region extrapolation and aids in processing chromosomes with large variations in boundaries.

Step 3: The estimate $\boldsymbol{e}_M$ is appended to $\boldsymbol{M(D)}$ at the $\boldsymbol{e}_1$ or $\boldsymbol{e}_2$ end for extrapolation in the upper or lower portion of D. The algorithm iterates through Steps 1 to 3 till $\boldsymbol{M(D)}$ extends through the length of the chromosome D.

3) Axis Smoothing and Geometric Correction

This step of the algorithm produces geometrically corrected or "straightened" chromosome D^2. Some preliminary processing is required before geometric correction. This is summarized below. Splines with knots at intervals of 3,4 and 8 points are fitted through $\boldsymbol{M(D)}$ successively to eliminate noise and provide a smoothened medial axis $\boldsymbol{M(D)}_S$. Then, $\boldsymbol{M(D)}_S$ is differentiated at every point with respect to x, so $\boldsymbol{M'(D)}_S$ is the vector describing the slope at each point (x, y) of D. Orthogonal lines $N(M)$ are calculated at each point on $\boldsymbol{M(D)}_S$ by,

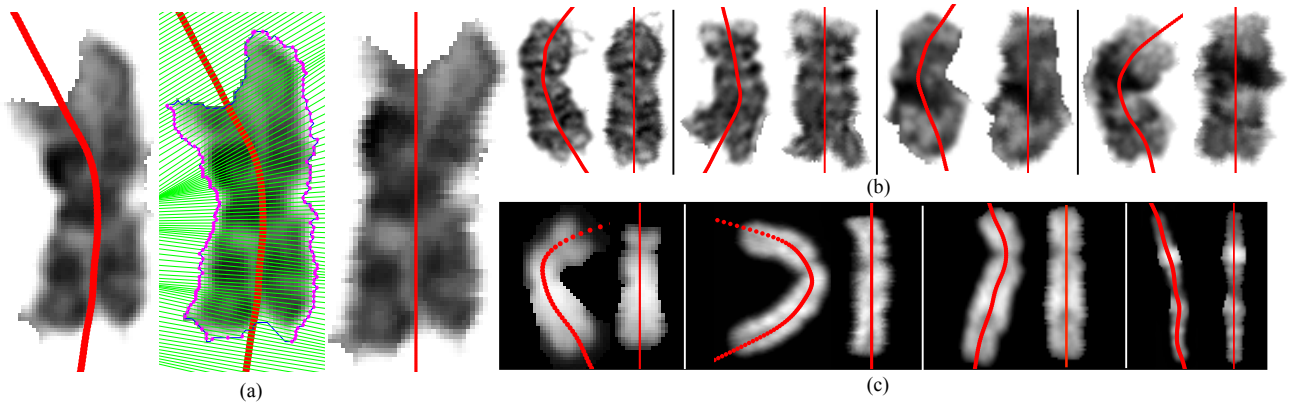


Figure 3. (a) A forked chromosomes that was geometrically corrected. (b) Examples of distorted and forked chromosome from LK_1 (c) The chromosomes with black background from Grisan *et al.* [4] data set that were geometrically corrected.

$$N_i(M_i) = -(M_i'(D))^{-1}{}_x + (M_{iy} - Mi'(D) \times M_{ix}) \quad (7)$$

Where $N_i(M_i)$ is the orthogonal line corresponding to the i^{th} point on $\boldsymbol{M}(D)_S$, Fig. 1 (g). Let $\boldsymbol{I}_{Ni}$ be the points of intersection of $N_i(M_i)$ with the original unsmooth chromosome boundary ∂D_0, Fig. 1 (g). Using this original boundary ensures that the parts of the chromosome which were eroded due to boundary smoothening are not lost during the geometrical correction. This further leads to more accurate feature extraction. For geometric correction a new destination image D^2 is created such that its width is twice the width of original chromosome D^1, Fig. 1 (g). The following discussion describes a chromosome as an image or matrix where d_{1ij} is the intensity value at the pixel belonging to i^{th} row and j^{th} column. Then, D^2 is populated as described below.

The profile ρ_i of the image between the two points of the $\boldsymbol{I}_{Ni}$ corresponding to i^{th} point on $\boldsymbol{M}(D)_S$ is obtained by connecting a straight path A_i of l points connecting the two points in $\boldsymbol{I}_{Ni}$. Here, l is the number of pixels in D that are traversed by A_i. The values in ρ_i are calculated by Nearest Neighbor Interpolation (NNI) method. Continuing this way, we obtain D^2, Fig. 1(i).

III. Results

This algorithm was tested on karyograms from LK_1 dataset. Fig. 3 shows few of the highly distorted and forked chromosomes that were geometrically corrected using our algorithm. The results show correspondence between the regions of deformed chromosome, and the regions of the geometrically corrected chromosome. Further, to test our algorithm's accuracy in revealing similarity between spatial distribution of intensity on chromosomes from the same class, band profiles of a pair of chromosomes from the same class was computed and has been shown in Fig. 2 (a). Additionally, we tested our algorithm on chromosomes from a high quality dataset from Grisan *et al.*, [4] the results of geometrical correction have been shown, Fig. 3 (c). Algorithm was found capable of extrapolating small seeds into medial axis spanning the entire chromosome, Fig. 2(b). Additionally, forked portion of the chromosomes were also recovered in the straightened chromosomes, Fig. 3 (a). The inclusion of a third parameter for extrapolation of seeds improved the geometrical correction results that we obtained previously. Thus, we were able to successfully correct the chromosomes that suffered from forking towards the ends, and correct the geometrical deformation that will help in more accurate feature extraction.

IV. Future Work

Using this algorithm, we have extracted features from chromosomes, which will be used for the classification process. We are working on the development of a robust classifier method to automatically karyotype the chromosomes from LK_1 dataset.

References

[1] A. Khmelinskii *et al.*, "A novel metric for bone marrow cells chromosome pairing," *IEEE Trans. on Biomed. Eng.,* vol. 57, pp. 1420-1429, 2010.

[2] J. Piper *et al.*,, "On fully automatic feature measurement for banded chromosome classification", Cytometry, vol. 10, pp. 242-255, 1989.

[3] J. Kao *et al.*, "Chromosome classification based on the band profile similarity along approximate medial axis," *Pattern Recognition*, vol. 41, pp. 77-89, 2008.

[4] E. Grisan *et al.*, "Automatic segmenta- tion of chromosomes in Q-band images," in *Proc. Annu. Int. Conf. IEEE EMBS*, 2007, pp. 5513–5516.

[5] J. Stanley *et al.*, "Datadriven homologue matching for chromosome identification," *IEEE Trans. Med. Imag.,* vol. 17, no. 3, pp. 451-462, June, 1998.

[6] X. Wu *et al.*, "Globally optimal classification and pairing of human chromosomes," in *Proc. 26th Annu. Int. Conf. IEEE EMBS*, Buenos Aires, 2010, pp. 2789– 2792.

[7] B. Lerner *et al.*, "A classification-driven partially occluded object segmentation (CPOOS) method with application to chromosome analysis," *IEEE Trans. Signal Process*., vol. 46, no. 10, pp. 2841–2847, Oct. 1998.

[8] X. Wang *et al.*, "A rule-based computer scheme for centromere identification and polarity assignment of metaphase chromosomes," *Comput.Methods Programs Biomed.*, vol. 89, no. 1, pp. 33–42, 2008.

[9] S. Khan *et al.*, "Robust band profile extraction using constrained nonparametric machine-learning technique," *IEEE Trans. Biomed. Eng*., vol. 57, no. 10, Oct. 2010.

[10] X. Bai, L. J. Latecki, andW. Y. Liu, "Skeleton pruning by contour parti- tioning with discrete curve evolution," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 29, no. 3, pp. 449–462, Mar. 2007.

[11] H. Blum, "Biological shape and visual science (part I)," *J. Theor. Biol.*, vol. 38, pp. 205–287, 1973.