IMAGE PROCESSING ALGORITHM FOR IDENTIFICATION OF CENTROMERE AND EXTRACTION OF FEATURES FOR AUTOMATED KARYOTYPING

Mousami V. Munot
Associate Prof. Dept. of E&TC, Pune Institute of Computer Technology, Pune, India,

Abstract: ***Classification of chromosomes in its classes is called as karyotyping. In the development of automated karyotyping schemes, the extraction and computation of chromosome image features as well as the selection and optimization of feature classifiers are two most important challenges. An optimal and small feature set is one of the important factors determining the performance and robustness of an automated classification system. Hence, a major research effort has gone into defining and searching for optimal features extracted from chromosome images. It is therefore necessary to devise an algorithm to derive the correct geometrical and the morphological features in cases of highly curved and bend chromosomes. This paper initially describes the traditional features used in the karyotyping of the chromosomes and further details effective image processing algorithms to extract the features of the curved and bend chromosomes. Various approaches and their experimental results are discussed. Profile density graphs of all 46 chromosomes, feature vector and the scatter plots of the features for a sample image in the database are also presented in this manuscript.***

Keywords: Chromosomes, karyotyping, classification, feature extraction

I. INTRODUCTION

A normal human cell has 22 pairs of autosomes and one pair of sex chromosomes. Analysis of chromosomes is extremely important when examining any genetic disorders in a human being [1]. Karyotyping is the process of classification of the chromosomes in its respective classes [2]. Various geometrical and morphological features reported in the literature for the process of chromosome classification. The G-banding profile was developed in the beginning of the 20th century and had become the most discriminative feature used in karyotyping. In the development of automated karyotyping schemes, various incremental learning algorithms [3, 4]the extraction and computation of chromosome image features [5] as well as the selection and optimization of feature classifiers are two most important challenges [6,7].Feature selection in developing a computer-assisted classification system can be regarded as a search, among all possible transformations (or extracted features), for the best subspace that preserves class separability as much as possible in the lowest possible dimensional space. An optimal and small feature set is one of the important factors determining the performance and robustness of an automated classification system. Hence, a major research effort has gone into defining and searching for optimal features extracted from chromosome images. Owing to the small size and limited resolution of banded chromosomes, finding effective features from original images is quite difficult [8, 9].

Various **Geometrical and morphological Features (Chromosome length, Area of chromosome, Perimeter of chromosomes, Number of bands above and below the centromere, width of the bands above and below the centromere, centromere Index CI, Band profile density)** The centromere forms the neck of the chromosomes. It connects the longer and the shorter arms of the chromosomes. Chromosomes have bands of high and low intensities (dark and light bands) along their major axis. Number of the bands (intensity variations) above and below the centromere is also an identified feature for automated karyotyping.

Accurate feature vector formed with efficient extraction algorithms may lead to (but not necessarily guarantee) improved classification accuracy and thus enhance the overall performance of AKS.

Features extracted and the algorithms used for extraction of the features play a substantial role in the performance of the developed AKS. A metaphase image imposes numerous challenges on the task of karyotyping. Some of these challenges include resolutions of the metaphase image, quality of the image, number of overlapping and touching chromosomes, bend and highly curved chromosomes [10]. Developing efficient algorithm to extract appropriate features irrespective of the challenges imposed demand great research efforts. Next sub-section details the features used in the development of AKS and the reported algorithms are detailed further. *Roshtkhari et al.* [10] presented novel algorithms for extrication of features in highly curved chromosomes. *Akila et al.* [11] proposed a hybrid algorithm but the effectiveness and success is limited to only groups A, B and C. *Jau et al.* [12] developed an efficient approach for medial axis determination and. *Shadab et al.[13]* and *Lerner Et al.* [14] developed an algorithm to estimate a single-line medial axis but algorithm failed in cases of severly bend chromosomes. Only 5 types of chromosomes were however classified as the part of the study. It is necessary to evaluate and confirm the importance of the features treating it as a 24 class problem. *Seung et al.* [15] selected some features: relative length, normalized Density Profile [DP] and *Gunter et al.* [16,17,18] proposed a novel approach exploiting the dominant points. Severely bent chromosomes are however not considered in this study. Shape variability in chromosomes is very natural due to its non-rigid nature. The efficiency of the algorithm in such cases needs to be further examined. The method formulated by *Enea et al.* [19] finally lead to good classification accuracies, whereas approach by but even these methods have limited success. Accurate feature vector formed with efficient extraction algorithms may lead to (but not necessarily guarantee) improved classification accuracy and thus enhance the overall performance of AKS, demanding further research in AKS.

II. PROPOSED METHODOLGY

Efficient image processing algorithm for the feature extraction of the chromosomes is highly desirous to improve the classification accuracy. A major challenge in feature extraction process is handling highly curved chromosomes to ensure accurate examination of the geometrical and morphological features. Popularly used approaches, based on Medial Axis Transform [MAT] prove to be inefficient in deriving the axis of the chromosomes as it ends up branching at the tapering ends of the chromosomes. This makes the extraction of the band profile density extremely challenging. The location of centromere in the highly curved chromosomes is another challenge in the feature extraction process. Wide assortment of the image processing algorithms has been developed to address the issue. Most of them are computationally complex and fail in case of highly curved chromosomes. This section utilizes the methods reported in the literature and with a careful study of ISCN diagram proposes a simple technique for the extraction of the features and location of the centromere in highly curved chromosomes.

A. Feature extraction (Average band intensity)

Non –rigid nature of chromosomes lead to occurance of unpredicatble shapes. It is therefore natural to encounter highly curved and bend chromosomes in the metaphase images. Proposed approach for feature extraction of curved chromosomes **:**

- Binarize the input image.
- Dilate the input chromosome image.
- Find the boundary of chromosome image.
- Find proper extended axis of chromosome.
- Do proper indexing of above obtained extended axis.
- Plot lines at interval of 10 degrees at every third point on obtained axis.
- Select line with minimum width for calculating average band intensity.
- Plot such bands on axis.
- Calculate average of each band on chromosome.
- Plot graph of 'Average Intensity Vs Band number'.

(a) (b)

Figure 1: Results of the feature extraction algorithm (straight and curved chromosomes)

Figure 1 depicts the experimental results to identify the average intensities along the longitudinal axis of highly curved chromosomes.

B. Automated Detection of Centromere in Chromosome

Automated identification of the centromere is obligatory to derive some of the geometrical and morphological features. This task is difficult in case of curved chromosomes.

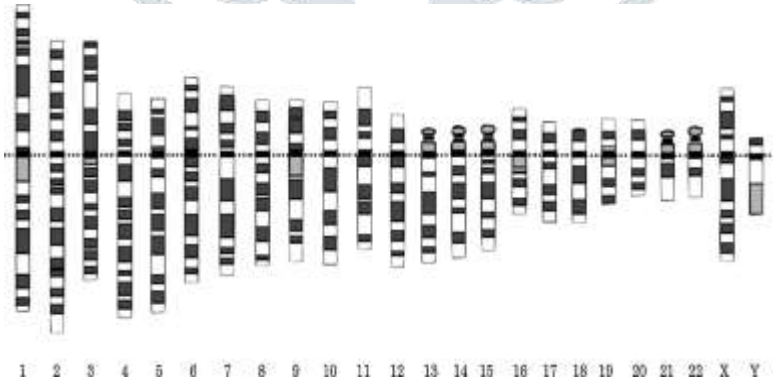


Figure 2: Ideograms of the chromosomes as per ISCN standard.

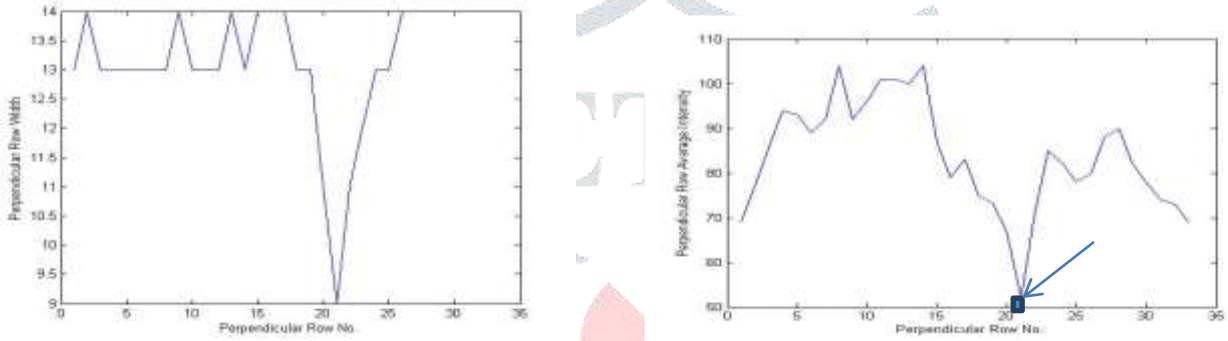
This section presents a novel approach based on careful study of ISCN standard. The two fundamental factors in the proposed approach are:

- The centromere forms the neck of the chromosomes. The width of the chromosome will therefore be relatively less in the area of the centromere. The perpendicular segment on the axis of chromosomes with smallest width will obviously be the centromere. It is necessary to mention that, when locating the centromere of larger chromosomes, the tapering portion of the chromosomes at the extreme ends is excluded to avoid the probability of centromere being identified on the extreme end of chromosome.
- A careful examination of the ideograms (band profile graphs) depicted in the ISCN standard reveal an important property of the centromere. Centromere (dotted line in Figure 2) of every chromosome is characterized by a darkest band, i.e. least intensity at centromere as seen in Figure 2.

Proposed Approach for automated identification of centromere

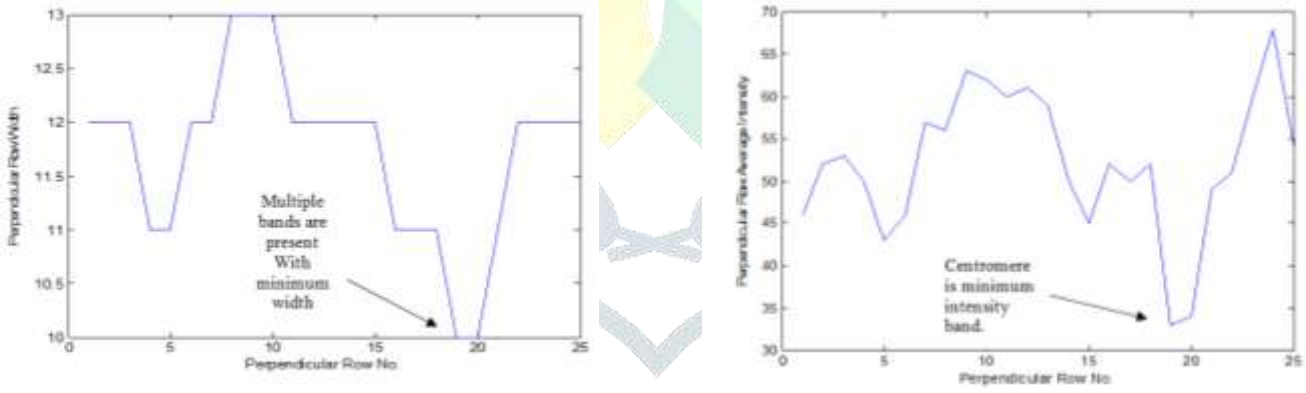
- Binarize the input image.
- Dilate the input chromosome image.
- Find the boundary of chromosome image.
- Find axis and plot bands.
- Calculate average of each band on chromosome.
- Plot graph of 'Average Intensity v/s Band number'.
- Plot graph of 'Band width v/s Band number'.
- Find centromere band as band with minimum intensity band.

Perpendicular segments are plotted on the axis of the chromosomes with the proposed algorithm. Figure 3(a) indicates a plot of perpendicular row no. v/s the corresponding width of the row. The centromere forms the neck of the chromosome. The perpendicular row with minimum width therefore signifies the position of the centromere. As clear in Figure 2, the ideograms of the chromosome indicate the presence of a dark band (minimum or very low intensity) at centromere. Figure 3(b) is a plot of average intensity at every perpendicular row. The minimum value of the intensity at row no. 21 conforms this observation. The developed approach successfully locates the centromere for the chromosomes in the data base. Some chromosomes show multiple minimum width bands. As seen from the Figure 4(a) and Figure 4(b), centromere is also the minimum intensity band of chromosome. In this case, the decision of the centromere position is done by multi-voting. When there are multiple minimum row width bands, the band with minimum intensity is chosen as centromere. A case with such a possibility is depicted in Figure 4.



(a) (b)

Figure 3:Estimation of Centromere using minimum intensity and width bands (a) Perpendicular row width Vs Band number (b) Average Intensity Vs Band number



(a) (b)

Figure **Error! No text of specified style in document.**: Identification of centromere in a case where multiple bands with least width are obtained.

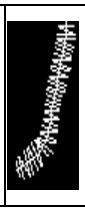
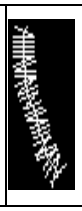
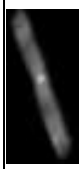
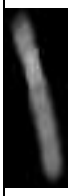
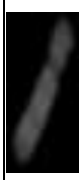
III. EXPERIMENTAL RESULTS

The plots the perpendicular segments on the axis of the chromosomes as discussed in the previous section. The perpendicular rows are plotted at an interval of every three pixels. Figure 5 below depicts the profile density graphs of all 46 chromosomes in a metaphase image. The average of the intensity values of the pixels along the perpendicular axis is plotted for every corresponding row. The chromosomes are polarized before plotting the profile density curves. The 22 groups of chromosomes are autosomes, forming homologous pairs. The band profile density of any pair of chromosomes must therefore be identical. This triggers the use of similarity based approaches for automated pairing of chromosomes. As clear from Figure 6, the practical images may not have exactly identical band patterns. The microscopic noise, staining effects, debris, blurred edges of the chromosomes, touching and overlapping possibilities lead to loss in the banding information and therefore makes the task of pairing the difficult. Table 1 describes the geometrical features of the same chromosomes. The length of the chromosomes in terms of the pixels is calculated. The Centromeric Index CI gives the ratio of the two arms of the chromosomes. The average intensity of the chromosomes on every perpendicular row is listed as a sequence of values. The last column of the table indicates the width of the chromosomes. The '0' in the sequence indicates the position of the centromere. The values on the either of the centromere signify the width of the intensity bands above and below the

chromosomes. The density plot is examined to initially identify the centromere; which is the least intensity value in the density plot and the perpendicular row of minimum width as explained earlier. A transition from the least value to high intensity value marks the presence of a light band on the corresponding chromosomes. The number of rows after which the transition takes place is identified as the width of the band . So *' -12, -8, 0, 2' indicates* presence of first light band above the centromere and the width of the band is of 8 rows (i.e. 8 X 3= 24 pixels). The next band in the sequence is identified using the next transition from high to low intensity, which is a dark band. The width of this band is 4 rows (-8 to -12). Similar logic applies on the lower side of the centromere. The width of the bands is marked with a positive value.

III. DISCUSSION AND CONCLUSION

This paper initially describes the traditional features used in the karyotyping of the chromosomes and further details effective image processing algorithms to extract the features of the curved and bend chromosomes. Various approaches and their experimental results are discussed. Profile density graphs of all 46 chromosomes, feature vector and the scatter plots of the features for a sample image in the database are also presented in this manuscript. Accurate feature vector formed with efficient extraction algorithms may lead to (but not necessarily guarantee) improved classification accuracy and thus enhance the overall performance of AKS. The change of chromosome band pattern resolution and boundary sharpness could affect the performance of this scheme. Potential negative influence of image resolution and quality, such as noise and band pattern fuzziness also has an impact on the algorithm An important point to be noted about the features is they are overlapping in nature. They are vague although within a specific range. There is randomness in their nature without any crisp boundary or specific threshold in the feature vector of the groups or the subgroups. Figure 7 details a scatter plot of two features namely: length of the chromosome and the CI. Figure.7(a) indicates the distribution of the features for Groups A and B and Figure 7(b) further details the distribution of the features for subgroups in group A. The distribution of the features signify the difficulty of the chromosome classification problem. The decision making about the chromosome class is therefore challenging. It requires expert knowledge and intelligence of the human expert to handle the fuzziness in the features . This makes the development of AKS a challenging task.

Chr.	Bands	Chr.	Bands	Chr.	Bands	Chr.	Bands	Chr.	Bands	Chr.	Bands	Chr.	Bands
1a		**1b**		**2a**		**2b**		**3a**		**3b**		**4a**	
													
4b		**5a**		**5b**		**6a**		**6b**		**7a**		**7b**	
													
8a		**8b**		**9a**		**9b**		**10a**		**10b**		**11a**	
													
11b		**12a**		**12b**		**13a**		**13b**		**14a**		**14b**	
15a		**15b**		**16a**		**16b**		**17a**		**17b**		**18a**	
18b		**19a**		**19b**		**20a**		**20b**		**21a**		**21b**	

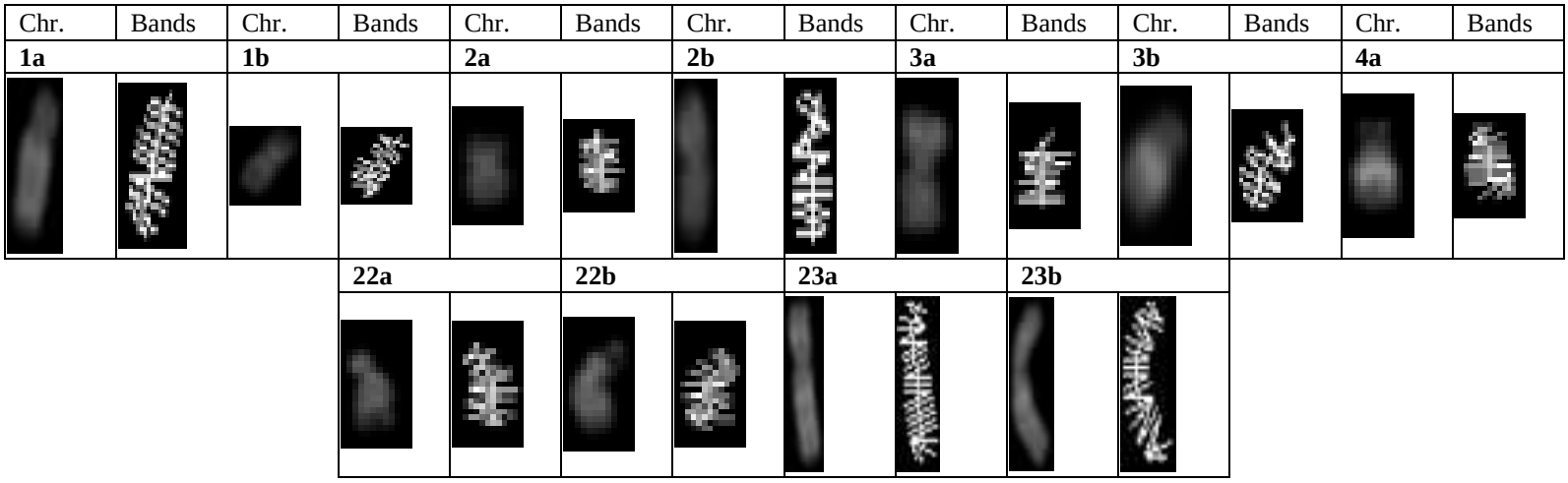
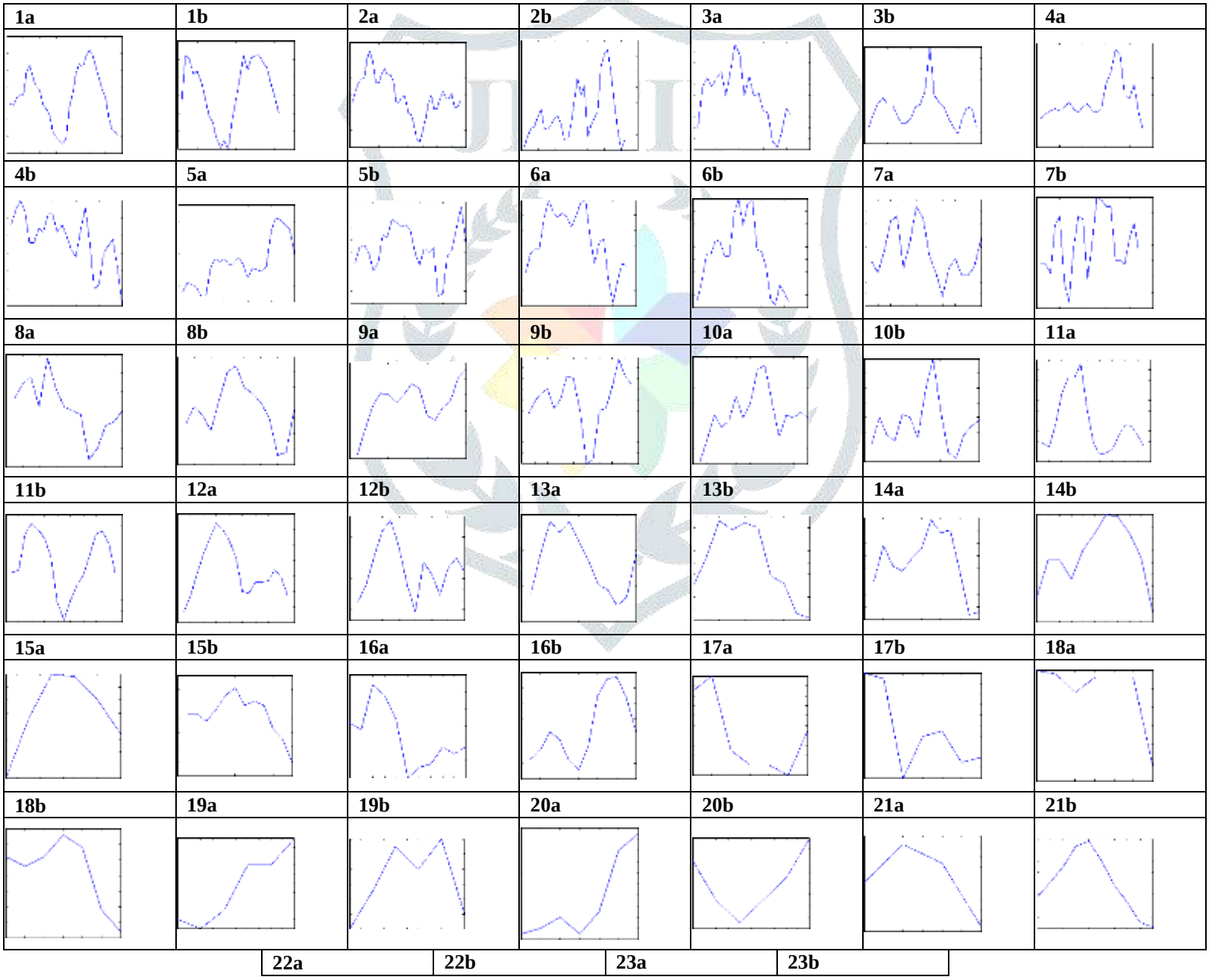


Figure 5. Chromosome and the axis with the perpendicular segments to derive the average intensity and other features of chromosomes



1a	1b	2a	2b	3a	3b	4a

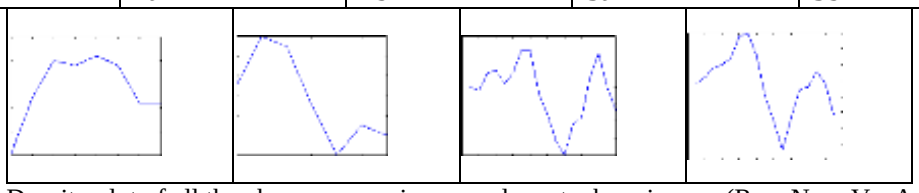
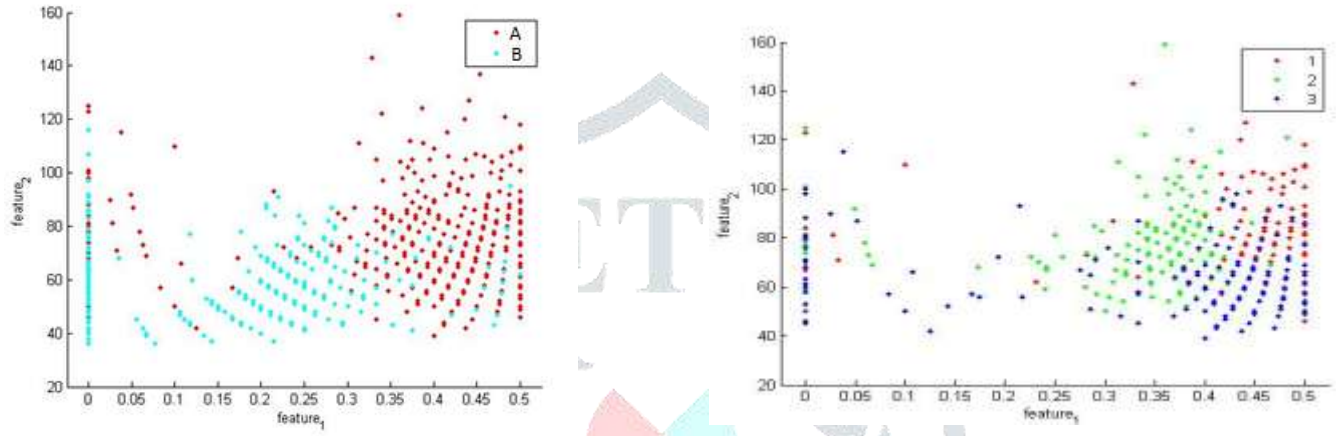


Figure 6: Band Profile Density plot of all the chromosomes in a sample metaphase image. (Row No; . Vs. Average intensity value)

Table 1. Geometrical features of all 46 chromosomes

Ch. No.	Len-gth	CI	Intensity	Distance between the bands
1a	78	0.38	59,79,38,78,84,88	-11,-7,0,3,8,9,11
1b	62	0.26	92,82,83,41,45,40,92,84,92	-5,-3,-2,0,4,5,6,10,11, 13
2a	86	0.36	85,67,72,75,71,68,56,49,39,46,60,52,52,58	-17,-15,-13,-12,-11,-9,-8,-3,0,1,2,4,5,6,7
2b	72	0.25	96,83,92,76,101,116,105,111,131,84	-2, 0 ,3 , 5 , 8 , 9 , 1 0 , 11 , 17,21
3a	51	0.42	98,94,101,90,114,90,99,90,91,66,84	-9,-8,-6,-5,-3,- 1,0,1,2,6,8
3b	58	0.41	93,77,87,74,59	-10,-7,-2,4,6
4a	55	0.2	92,91,96,90,95,90,126,98,106	-19,-18,-16,-14,-12,-10,-6,-3,-2
4b	59	0.2	95,83,83,87,86,91,91,85,79,93,80	-17,-15,-14,-13,-12,-11,-10,-7,-5,-3,1
5a	60	0.04	62,54,71,77,75,77,74,73,65,70,94	1,4,6,7,8,9,11,13,14,16,19
5b	59	0.24	53,43,57,56,63,60,61,45,52,50,52,33,68	-16,-14,-12,-11,-10,-8,-7,-4,-3,-2,-1,0,5
6a	56	0.34	74,76,74,74,79,68,66,54,53	-10,-8,-7,-5,-2,0,2,4,6
6b	52	0.42	87,92,86,104,99,82,68	-5,-3,-1,0,2,7,8
7a	45	0.38	69,93,71,97,59,75,68	-5,-2,-1,1,5,7,8
7b	53	0.5	64,76,58,69,78,71,82,65	-9,-7,-5,-3,-1,1,5,7
8a	37	0.28	64,56,69,42	-7,-6,-5,0
8b	37	0.07	74,66,87,58	-11,-9,-6,-1
9a	40	0.33	81,75,87,64	-6,-4,-2,1
9b	43	0.47	60,51,66,25,74	-5,-4,-2,1,6
10a	42	0.25	56,52,62,55,72,49,56,55,57	-9,-8,-6,-5,-2,0,1,2,3
10b	40	0.26	70,62,71,63,90,56	-9,-7,-6,-4,-2,1
11a	43	0.41	65,132,68,76	-5,-1,2,6
11b	43	0.35	81,36,77	-7,-2,4
12a	44	0.05	114,69,77	5,10,13
12b	38	0.28	83,39,63,47,65	-5,-2,-1,1,3
13a	34	0	126,120,126,79	-9,-8,-7,-2
13b	30	0	98,94,97	-7,-6,-5
14a	33	0.16	64,54,74,69,70,37	-8,-6,-3,-2,-1,1
14b	32	0.09	81,81,92	-8,-7,-4
15a	22	0.33	115	1
15b	33	0	53,66,71,59	-10,-7,-6,-5
16a	32	0.36	59,72,45,54,52	-5,-4,-1,2,3
16b	34	0.5	81,56,117	-3,0,4
17a	24	0	68,50	-5,-2
17b	24	0.42	50,59,53	-1,1,2
18a	24	0	79,83	-4,-3
18b	24	0	68,78	-5,-3
19a	22	0.333	39	0
19b	22	0.33	59,56,60	-1,0,1

Ch. No.	Len-gth	CI	Intensity	Distance between the bands
20a	23	0.14	54,51	2,3
20b	21	0.5	43	0
21a	18	0	91	-2
21b	29	0.01	108	-5
22a	25	0.12	65,64,66	2,3,4
22b	23	0.14	72, 60,63	-4 -1 0
23a	45	0.33	79,85,81,91,64,83	-10,-8,-7,-5,-1,3
23b	43	0.23	85,48,73	-6,-2,2



(a) (b)

Figure 7. Scatter plot of the features. Distribution of the features: chromosome length and CI (a) plotted for group A and B, (b) plotted for subgroups 1,2 and 3 in group A.

REFERENCES:

[1] Andrew Carothers and Jim Piper, "Computer-aided classification of human chromosomes: a review", *Statistics and Computing 4*, pp 161-171, 1994.

[2] Nina Sweeneyt, Robert L. Beckert, and Brian Sweeney, "A Comparison of Wavelet and Fourier Descriptors for a Neural Network Chromosome Classifier", *International Conference of Engineering in Medicine and Biology Society*, vol.3, pp 1359-1362, 1997.

[3] Prachi Joshi, Mousami Munot, Parag Kulkarni, Madhuri Joshi, "Efficient Karyotyping of Metaphase Chromosomes Using Incremental Learning", The IET Journal of Science Measurement and Technology, Vol.7, Issue 5, September 2013, pp. 287-295.

[4] Mousami Munot, Prachi Joshi, Madhuri Joshi, Parag Kulkarni, "An Incremental Approach for Efficient Karyotyping Systems", Journal of Medical Imaging and Health Informatics, Med. Imaging Health, Inf. 6, Febraury 2016, pp. 221-225.

[5] Mousami Munot, Jayanta Mukhopadhyay, Madhuri Joshi, "A Novel Approach for Efficient Extrication of Overlapping Chromosomes in Automated Karyotyping", Springer Journal of Medical and Biological Engineering and Computing, August 2013, pp. 1-14.

[6] Mousami Munot, Prachi Joshi, Madhuri Joshi, Parag Kulkarni, "Efficient Pairing of Chromosomes in Metaphase Image for Automated Karyotyping", IEEE- EMBS Conference on Biomedical Engineering and Sciences, IECBES - 2012, Langkawi, Malaysia, Dec-2012,pp:916-921.

[7] Mousami Munot, Madhuri Joshi, "Towards Automated Karyotyping of Curved Chromosomes", Coimbatore Institute of Information Technology,CiiT International Journal of Artificial Intelligent Systems and Machine Learning, January 2012.

[8] Mehdi Moradi, S. KamaledinSetarehdan, "New features for automatic classification of human chromosomes: A feasibility study", *Pattern Recognition Letters 27*, pp 19–28, 2006.

[9] ArtemKhmelinskii, Rodrigo Ventura and JoãoSanches, "Chromosome Pairing For Karyotyping Purposes Using Mutual Information", *5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp 484-487, 2008.

[10] MehrsanJavanRoshtkhari and SeyedKamaledinSetarehdan, "A novel algorithm for straightening highly curved images of human chromosome", *Pattern Recognition Letters 29*, pp 1208-1217, 2008.

[11] AkilaSubasingheArachchige, JagathSamarabandu, Joan Knoll, Wahab Khan, Peter Rogan, "An Image Processing Algorithm for Accurate Extraction of the Centerline from Human Metaphase Chromosomes", *IEEE International Conference on Image Processing*, pp 3613-3616, September 2010.

[12] Jau-hong Kao, Jen-hui Chuang, Tsaipei Wang, "Chromosome classification based on the band profile similarity along approximate medial axis", *Pattern Recognition 41*, pp 77 – 89, 2008.

[13] Shadab Khan, Joao Sanches and Rodrigo Ventura, "Robust Band Profile Extraction Using Constrained Nonparametric Machine-Learning Technique", *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 10, pp 2587-2591, October 2010.

[14] Lerner B., Rosenberg B., Levinstein M., Guterman H., Dinstein I. and Romem Y., "Medial Axis Transform Based Features and a Neural

Network for Human Chromosome Classification". Elsevier Journal on Pattern Recognition, vol. 28, issue 11, pp 1673-1683, November 1995.

[15] Seung Yun Ryu, Jong Man Cho, SeungHyo Woo, "A Study for the Feature Selection to Identify Giemsa-Stained Human Chromosomes Based on Artificial Neural Network", *IEEE International Conference on EMBS*, Istanbul, Turkey, pp 691-692, 25-28 October 2001.

[16] Gunter Ritter, GernotSchreib,"Using dominant points and variants for profile extraction from chromosomes", *Pattern Recognition 34*, pp 923-938, 2001.

[17] Gunter Ritter and GernotSchreib, "Profile and feature extraction from chromosomes", *IEEE Conference on Pattern Recognition*, vol.2, pp 287-290, 2000.

[18] Gunter Ritter, ChristophPesch, "Polarity-free automatic classi_cation of chromosomes", *Elsevier Science of Computational Statistics & Data Analysis 35*, pp 351-372, 2001.

[19] EneaPoletti, Enrico Grisan, and Alfredo Ruggeri, "Automatic classification of chromosomes in Q-band images", *IEEE Conference on Engineering in Medicine and Biology Society (EMBS)*, pp 1911-1914, 2008.

[20] Xingwei Wang, Bin Zheng, Shibo Li, John J. Mulvihill, Hong Liu, "A rule-based computer scheme for centromere identification and polarity assignment of metaphase chromosomes", Elsevier *Journal of computer methods and programs in biomedicine 89*, pp 33–42, 2008.

[21] L. V. Guimaraes, A. Schuck Jr and A. Elbern, "Chromosome Classification for Karyotype Composing applying Shape Representation on Wavelet Packet Transform", *International Conference of the IEEE EMBS*, Vol.1, pp 941-943, September 17-21, 2003.

[22] Qiang Wu; Castleman, K.R., "Automated chromosome classification using wavelet-based band pattern descriptors", *IEEE Conference on Computer-Based Medical Systems*, pp 189-194, 2000.

[23] Jim Piper and Erik Granum, "On Fully Automatic Feature Measurement for Banded Chromosome Classification", *Alan R. Liss, Inc. Cytometry 10*, pp 242-255, 1989.

[24] Lerner, B., Levinstein, M., Rosenberg, B., Guterman, H., Dinstein, I. and Romem, Y, "Feature Selection and Chromosome Classification Using a Multilayer Perceptron Neural Network", *IEEE World Congress on Computational Intelligence., IEEE International Conference on Neural Networks*, vol.6, pp 3540–3545, 1994.