Feature Extraction for the Classification of Human Chromosomes from G-Band Images using Wavelets

R.Nandakumar[1] and K.B.Jayanthi[2]
[1]Electronics and Communication Engineering, K S R Institute for Engineering and Technology
[2]Electronics and Communication Engineering, K.S.Rangasamy College of Technology

Abstract - **Chromosomes contain genes that provide the coded information for human beings to grow, develop and function. Any change in the number, size or structure of the chromosome leads to chromosomal abnormalities which will cause birth defects. However many of these defects are preventable, if detected earlier. The main objective of this work is to determine good features to classify human chromosomes and to detect chromosomal abnormalities from G-Band chromosome images. As a first step, chromosome images are analyzed using Discrete Wavelet Transform (DWT) to get coefficients which contain information about the banding pattern. The banding patterns allow a chromosome to be reliably differentiated from other chromosomes of same size and centromere position. From the coefficients, statistical features are calculated. A neural network may be used for further classification using these features.**

Index Terms- **Chromosome, DWT, Statistical Features, Neural Networks, Chromosomal abnormalities**

I. INTRODUCTION

Globally, at least 7.6 million children are born annually with severe genetic or congenital malformations. 90 % of these are born in mid and low income countries. The genetic and congenital disorder is the second most common cause of infant and childhood mortality and occurs with a prevalence of 25-60 per 1000 births. Genetic diseases can vary in severity, from being fatal before birth to requiring continuous management; their onset covers all life stages from infancy to old age [1]. In India 25 million births occur annually. Chromosomal abnormalities form a major part of genetic disease burden in India. Nearly 5,00,000 babies are born with some form of birth defects every year. It occurs when an individual is affected by a change in the number, size or structure of his or her chromosomes.

Down syndrome is the most common genetic disorder caused by a chromosomal abnormality. It affects 1 out of every 800 to 1,000 babies. It occurs when some or all of a person's cells have an extra full or partial copy of chromosome 21.The most common form of Down syndrome is known as Trisomy 21. Individuals with Trisomy 21 have 47 chromosomes instead of the usual 46 in each of their cells.Only two other trisomies have been observed in babies born alive (trisomies 13 and 18), but babies born with these trisomies have only a 5% chance of surviving longer than one year[1]. Chromosomal abnormality is an important cause of mental retardation. Apart from sex specific genes present on X and Y chromosomes some autosomal genes also play a role in sex determination. Any alteration in the genes, gene dosage or the sex chromosomes lead to abnormalities of sexual development, ranging from complete sex reversal to hermaphroditism [2].

Many genetic disorders or possible abnormalities that may occur in future generations can be predicted by analyzing the shape and morphological characteristics of chromosomes[3]. Automated chromosome classification is an essential task in cytogenetics and has been an important pattern recognition problem. Numerous attempts are being made to characterize chromosomes for the purpose of clinical and cancer cytogenetic research.

It is important to determine good features for chromosome classification. Centromere intensities are believed to be important differentiating features of homologous chromosomes[4]. Each chromosome displays a unique banding pattern. Specific pairs of chromosomes can be identified using the centromere position and arm ratios. Inevitably several pairs of chromosomes appear identical by these criteria and some of the chromosomes are also overlapped. Many algorithms have been tried in the past to separate the touching chromosomes and overlapping chromosomes[5],[6].

The concept of automated chromosome classification has been under research for many years [7],[8]. Commercial systems available now are very costly. It is not possible for common people to make use of this facility in developing countries. Moreover there is a scope to improve the recognition rate and accuracy [9]. Hence an effort is made in this work to develop a cost effective automated classifier for chromosomes using G-Band chromosome images.

G-Band is the most common type of banding. It generates a distinct transverse banding pattern characteristic of each class which is an important feature for chromosome classification and pairing The idea of using G-band images is that, it helps to identify even tiny abnormalities [10].

II. SPATIAL DOMAIN ANALYSIS

A. Length of Chromosome

Centromere divides the chromosome into two 'arms': the short arm, called the `p' arm, and a long arm called the `q' arm. The relative position of the centromere is constant, which means that the ratio of the lengths of the two arms is constant for each chromosome. This ratio is an important parameter for chromosome identification, and also, the ratio of lengths of the two arms allows classification of chromosomes. The length decreases from chromosome 1 to 23. But some of the chromosomes have same length and variation is also very less.

B. Properties of Chromosome Image

For a bending chromosome, it is very difficult to measure the length unless it is straightened. So, the image is converted as a binary image as shown in the Fig. 1

Fig. 1. Input image and Binary image

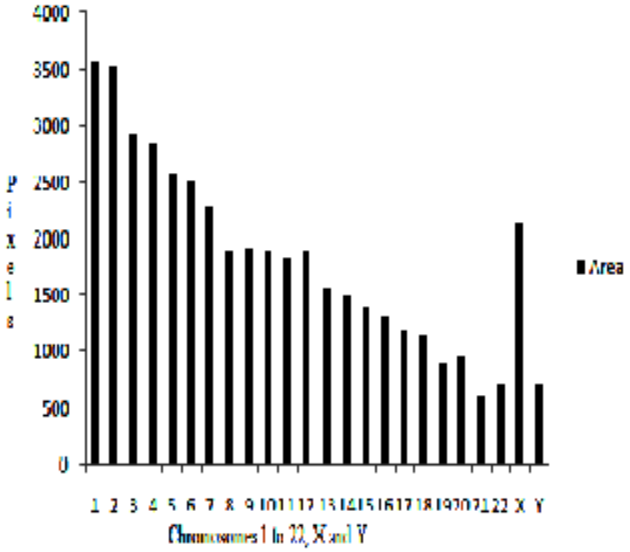

Fig. 2. Comparison of area

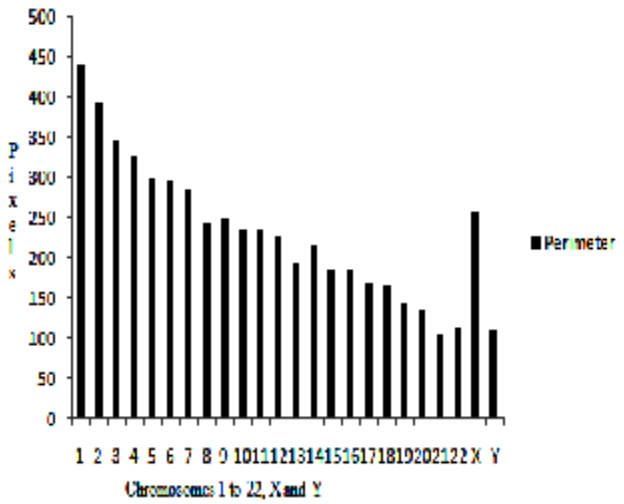

Fig. 3. Comparison of perimeter

Properties like area and perimeter are calculated for the chromosomes 1 to 22, X and Y and the comparison is given in Fig. 2 and Fig. 3 respectively. It is observed from the comparison that both values of area and perimeter are decreasing for the autosomes 1 to 22. But still some of the chromosomes have same values. The values of sex chromosomes X and Y are overlapping with the autosomes.

C. Banding Pattern

It is clear from the comparison, that properties alone cannot be used to classify the chromosomes. So, the banding pattern of chromosomes is analysed. Fig. 4 shows the intensity profile of the banding pattern of a straight chromosome. Fig. 5 shows the intensity profile of the banding pattern of a bending chromosome. The intensity profile gives the variation of gray levels with respect to the pixels along the axis in spatial domain. Each peak in the graph represents the corresponding white band and each valley represents the corresponding black band of the chromosome image.

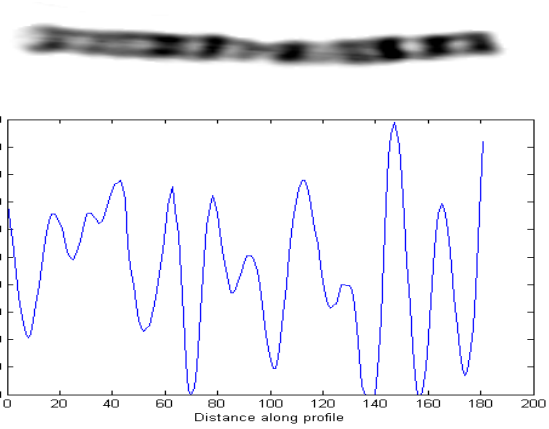

Fig. 4. Intensity profile of a straight chromosome

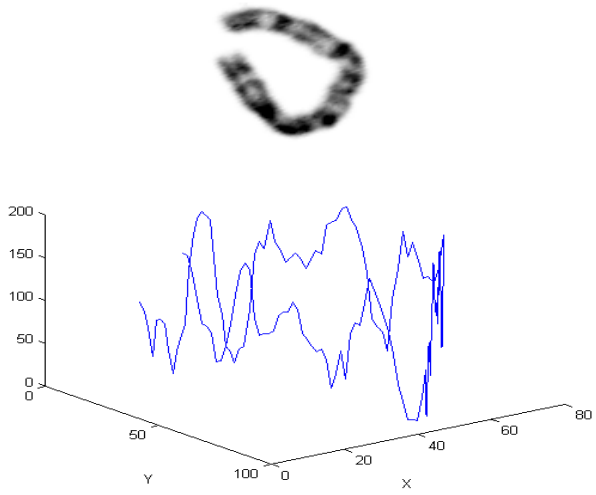

Fig. 5. Intensity profile of a bending chromosome

The banding pattern can be analyzed effectively for a straight chromosome using this intensity profile. In case of a bending chromosome, analysis of banding pattern using the intensity profile is complicated in the spatial domain as shown in Fig. 5. Therefore, in this work banding pattern is analysed in the transform domain using Discret Wavelet Transform(DWT).

III. PROPOSED SYSTEM

The process of extracting features from the chromosome images for classification process is automated by the proposed system shown in Fig. 6. The extracted features may be given to a classifier for further classification using a neural network [11].

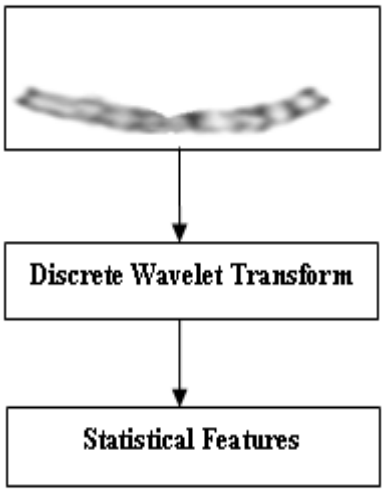

Fig. 6. Wavelet based classifier

A. Preprocessing

Chromosome bands are not clearly visible in the input chromosome images. In preprocessing, contrast of the chromosome image is improved to differentiate the bands clearly. First the color image is converted to a grayscale image. Histogram equalization and other enhancement techniques are tried on the input image to increase the contrast. These methods remove the banding pattern which is the unique feature to identify a particular chromosome. So contrast of the image is increased without affecting the banding pattern.

B. Discrete Wavelet Transform

Unlike the Fourier transform, whose basis functions are sinusoids, wavelet functions are based on small waves, called wavelets, of varying frequency and limited duration. In the previous works Discrete Cosine Transform is used for boundary mapping and classification of chromosome spread images [12]. In DCT, the basis function is fixed, where as in DWT different basis functions are available depending upon the choice of the wavelet. DWT can be applied to an entire image without imposing block structure as used by the DCT, thereby reducing blocking artifact. This allows them to provide localization in both the spatial and frequency domain. Also multiresolution is possible with DWT. Thus DWT gives higher flexibility than DCT and provides better identification of which data is relevant to human perception. In this work Discrete Wavelet Transform is used to analyze the chromosome images in the frequency domain and to compress the frequency components [13].]. It decomposes an image into a set of different resolution sub-images, corresponding to the various frequency bands. DWT of an image f(x,y) of size MxN is given by

$$W\varphi(j_0,m,n)=\frac{1}{\sqrt{MN}}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}f(x,y)\varphi_{j_0,m,n}(x,y) \quad (1)$$

$$W_\psi^i(j,m,n)=\frac{1}{\sqrt{MN}}\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}f(x,y)\psi_{j,m,n}^i(x,y) \quad (2)$$

where

$i = \{ H,V,D\}$

$W_\varphi(j_0,m,n)$ - coefficients define an approximation of f(x,y) at scale j_0

$W_\psi(j,m,n)$ – coefficients add horizontal(H), vertical(V) and diagonal(D) details for scales $j \geq j_0$.

At first, the input chromosome image is analyzed using Symlets, short for symmetrical wavelets and Daubechies wavelets and the analysis is to be continued with other wavelets to identify the best. Symlets are designed to have the least asymmetry and

highest number of vanishing moments for a given compact support which is the interval in which the function has non zero values. This compactly supported nature enables temporal localization of features. Daubechies wavelets are a family of orthonormal, compactly supported scaling and wavelet functions that have maximum regularity. The image is decomposed into four bands namely Wφ- coefficients of approximation (LL), W_ψ^H- coefficients of horizontal details(LH), W_ψ^V- coefficients of vertical details(HL) and W_ψ^D- coefficients of diagonal details (HH) using equations (1) and (2). The first level decomposition of chromosome 1 using these wavelets is shown in the Fig. 7

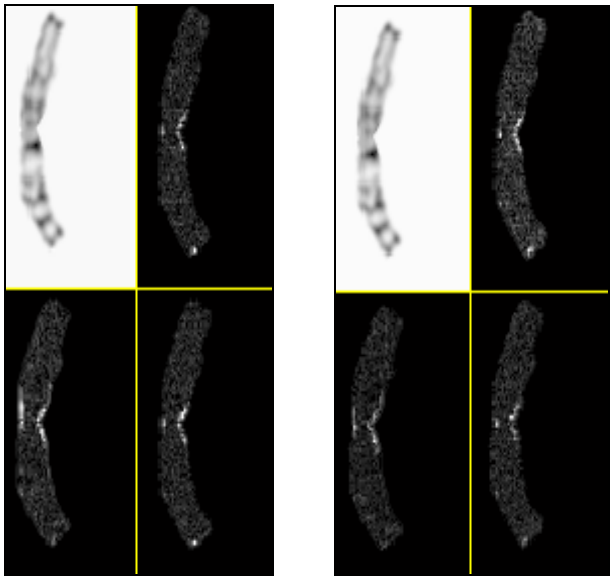

Fig. 7. Decomposition using symlets and daubechies

C. Statistical Features

Using the coefficients of first level decomposition, statistical features Mean, Mode and Standard Deviation (SD) are calculated using the equations (3), (4) and (5).

$$I_{mean}(k,l)=\frac{1}{N_R}\sum_{m}\sum_{m\in\Re}\left[u(m-k,n-l)\right] \quad (3)$$

$$I_{mode}=I(k,l)of\ \max\left[h(x_i)\right] \quad (4)$$

$$I_{SD}(k,l)=\frac{1}{N_R}\sum_{m}\sum_{m\in\Re}\left[u(m-k,n-l)-M_1(k,l)\right] \quad (5)$$

The features are calculated using the coefficients of approximation, horizontal, vertical and diagonal details of both first level and second level decompositions. By analyzing the features, it is observed that the values of the features of approximation get multiplied by 2 for each level and the features of details are very less or negative values. Statistical features of the coefficients of approximation W_φ are taken for further analysis as most of the information is contained in that. These features can be given as inputs to the classifier for classifying and pairing chromosomes [14], [15].

TABLE 1
STATISTICAL FEATURES OF CHROMOSOME 1

Sample	db4			sym4		
	Mean	Mode	SD	Mean	Mode	SD
1	488.2	511.1	59.36	488.2	508.5	59.35
2	483.8	513.5	69.51	483.8	513.5	69.49
3	486.2	512.6	70.73	486.2	506.3	70.73
4	488.5	507.3	63.77	488.6	506.4	63.78
5	493.3	509.1	58.03	493.3	508	58.01
6	495.3	508	50.36	495.3	508.2	50.37
7	487.7	509.2	69.58	487.7	509.7	69.58
8	487.3	512.3	68.92	487.3	509.9	68.92
9	495.2	513.7	52.66	495.3	512.2	52.66
10	494.6	510.3	53.33	494.6	513.7	53.33
11	497.2	509.2	48.14	497.2	508.6	52.09
12	496.8	508.8	49.22	496.9	510.6	49.21
13	489.3	514.1	57.92	489.3	511.4	57.91
14	484.9	508	66.09	484.9	505.9	66.11
15	487.4	507.4	60.24	487.4	508.3	60.24
16	489.8	506.3	70.01	489.8	515.1	70.01
17	491.1	506.5	63.56	491.1	507.5	63.56
18	492.8	512.1	68.96	492.8	507.3	68.96
19	492.4	511.9	63.16	492.5	515.9	63.16
20	491	510.5	65.39	491	510.1	65.39
21	490	508.5	58.59	490	512.3	58.59
22	484.1	512.7	68.77	484.1	510.2	68.78
23	485.1	509.5	65.03	485.1	510.4	65.03
24	487.5	510.6	64.01	487.5	511.8	64.01
25	490.3	512	58.67	490.3	511.2	58.67
26	488.9	510	62.89	488.9	506.1	62.91
27	490.1	511.6	65.65	490.1	513.9	65.65
28	494.8	514.9	62.17	494.8	513.5	62.17
29	496	508.2	53.98	496	510.4	53.98
30	494.6	510.7	60.59	494.6	511.3	60.6

IV. RESULTS

Input images are prepared from the karyotyped G band chromosome images of size 768 x 576. 30 samples of chromosome 1 and a set of images for chromosome 1 to 22, X and Y are taken for analysis. After preprocessing, the chromosme image is divided into subbands of coefficients using Symlets and Daubechies wavelets. A set of statistical features Mean, Mode and SD obtained using Symlet (sym4) and Daubechies wavelets (db4) for 30 samples of chromosome 1 is given in the Table. 1 Comparison of features obtained using symlets and daubechies wavelets is shown in the Fig. 8 and Fig. 9 respectively.

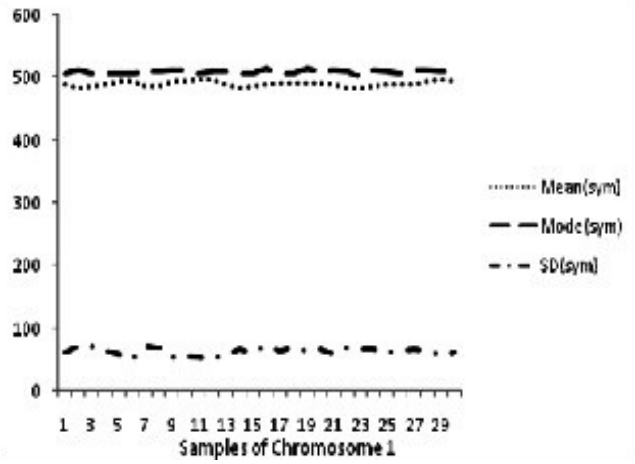

Fig. 8. Features of chromosome 1(Symlet)

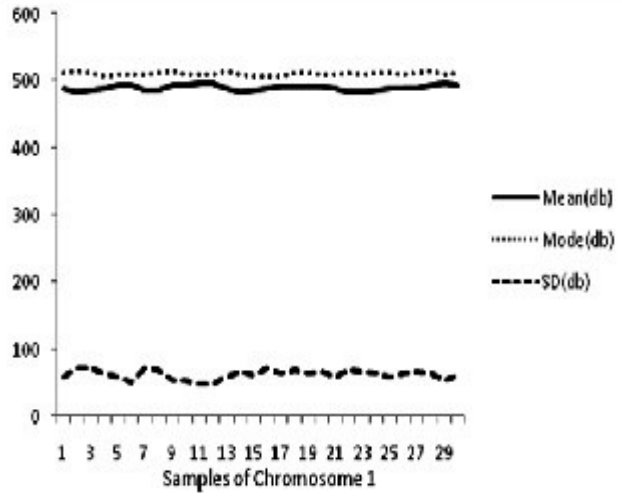

Fig. 9. Features of chromosome 1(Daubechies)

It is found from the Table 1, Fig. 8 and Fig. 9, that the features mean and SD are same for both the wavelets, but the values of mode are different with a marginal difference. Symlets give close values of mode for equal values of mean than Daubechies. The values of the features mean and mode lie in a narrow band for all 30 samples of chromosome 1and they can be used to classify chromosome 1 from others. The observations are seen true from the Fig. 10 where the features obtained using Symlets and Daubechies have overlapped.

Another set of chromosomes 1 to 22, X and Y taken from the same karyotyped image are analysed using the Symlets and the statistical features are calculated. Variation in the values of mean and mode between the 22 autosomes and sex chromosomes X and Y is shown in the Fig.11. Vaiation in the Standard Deviation(SD) is shown in the Fig. 12.

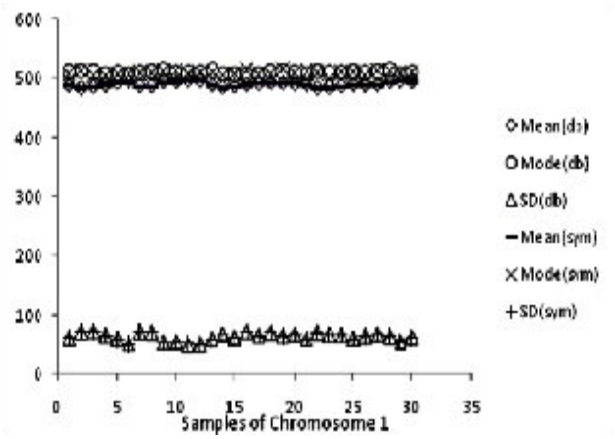

Fig. 10. Features of chromosome 1(Symlets & Daubechies)

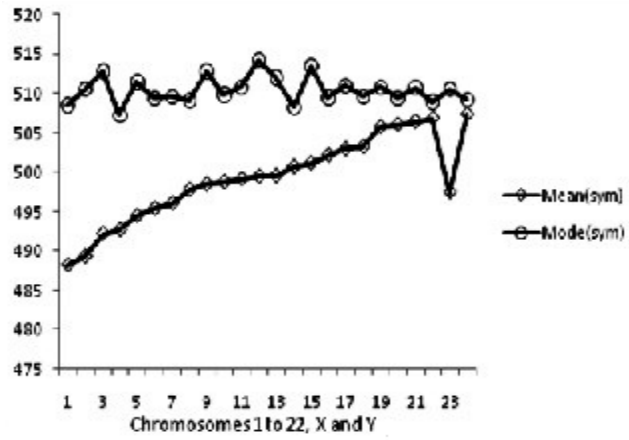

Fig. 11. Mean and Mode values of chromosomes 1 to 22, X & Y

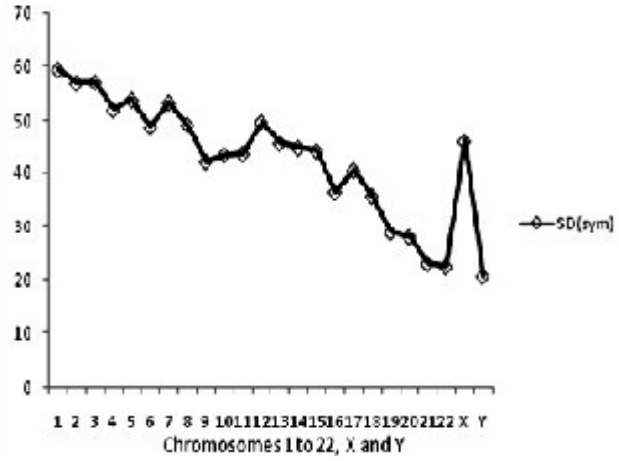

Fig. 12. SD values for chromosomes 1 to 22, X and Y

It is clear from the Fig. 11 and Fig. 12 that the value of mean is increasing continuously and the value of SD is decreasing with the chromosome number for the autosomes 1 to 22. The valley and peak points are due to X chromosome, as its features overlap with the autosomes.

V. CONCLUSION

A method is proposed to extract features for classification of the human chromosomes from G band metaphase images using Discrete Wavelet Transform. Initially chromosome images are converted to binary images in the spatial domain and features like area and perimeter are measured. As several chromosomes appear identical with these features and banding pattern is completely removed in the binary image, it is insufficient to classify the chromosomes. Banding pattern is analyzed again by plotting the intensity profile. But that is complicated for bending chromosomes. So the banding pattern which is a unique feature of chromosomes is taken to transform domain and the frequency components are analyzed by applying Discrete Wavelet Transform using Symmetrical and Daubechies wavelets. Statistical features mean, mode and standard deviation (SD) are calculated from the coefficients of approximation. Features are analyzed and compared for different sets of chromosome images. The mode values do not show any significant difference between the different chromosomes and hence it cannot be used as a good feature to classify the chromosomes.The features mean and SD can be used further for classification as they show good variations between the chromosomes.

Future work will focus on further testing of the proposed method in a larger image dataset to validate the features and analyze the chromosome images using other wavelets, wavelet packets and multi wavelets to determine new features for classification of human chromosomes.

ACKNOWLEDGEMENT

The authors would like to thank Dr. Suresh from Mediscan systems, Chennai, India for having kindly provided the inputs for this work.

REFERENCES

[1] Anupam Kaur and Jai Rup Singh, "Chromosomal abnormalities:Genetic Disease Burden in India", Int J Hum Genet, 10(1-3): 1-14(2010).
[2] Damiani D, Fellous M, Mcelreavy K, " True hermaphroditism: Clinical aspects and molecular studies in 16 cases", Eur J Endocrin, 136: 201-204.
[3] Charles C.Tseng. Human chromosome analysis. 33-56, 1995.
[4] Parvin Mousavi, Rabab Kreidieh Ward and Peter M.Lansdorp, "Feature Analysis and Centromere Segmentation of Human Chromosome Images Using an Iterative Fuzzy Algorithm", IEEE Transactions on Biomedical Engineering, Vol.49, No.4, pp 363-371, April 2002.
[5] R.J.Stanly, J.Keller,P.gader and C.V.Caldwell, "Homologue Matching applications: Recognition of overlapped Chromosomes",Pattern Analysis and Application, pp 206-217, 1998.
[6] Enrico Grisan, Enea Poletti and Alfredo Ruggeri, "Automatic Segmentaion and Disentangling of Chromosomes in Q-Band Prometaphase images", IEEE Transactions on Information Technology in Biomedicine, Vol.13, No.4, pp 575-581, July 2009.
[7] Enea Poletti, Enrico Grisan and Alfredo Ruggeri, "Automatic classification of chromosomes in Q-band images", International IEEE EMBS Conference Vancouver, British Columbia, Canada, pp 1911-1914,August 2008.
[8] Benoit Legrand, Che Sau Chang, Sim-Heng ong, Soek-Ying Neo and Nallasivam Palanisamy, "Automated Identification of Chromosome Segments Involved in Translocations by Combining Spectral Karyotyping and Banding Analysis", IEEE Transactions on Systems, MAN and Cybernetics-Part A: Systems and Humans, Vol.38,No.6, pp 1374-1383 November 2008.
[9] Mehdi Moradi and Kamaledin, " New features for automatic classification of human chromosomes: A feasible study", Pattern Recogniition Letters 27 (2006) 19-28.
[10] Rodrigo Ventura, Artem Khmelinskii, "Classifier-assisted metric for chromosome pairing", 32nd Annual International Conference of the EMBS, pp 6729 – 6732, August 2010.
[11] Boaz 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 244-252, August 1998.
[12] Qiang Wu, Zhongmin Liu, Tiehan Chen, Zixiang Xiong and Kenneth R. Castleman, " Subspace- Based Prototyping and Classification of Chromosome Images", IEEE Transactions on Image Processing, Vol.14, No.9, pp 1277-1287 September 2005.
[13] Zhongmin Liu, Zixiang Xiong, Qiang Wu, Yu-ping Wang and Kenneth R. Castleman, " Cascaded Differential and Wavelet Compression of Chromosome Images", IEEE Transactions on Biomedical Engineering, Vol.49, No.4, pp 372-383, April 2002.
[14] Ibrahiem M. M. El Emary, "On the Application of Artificial Neural Networks in Analyzing and Classifying the Human Chromosomes", Journal of Computer Science 2 (1): pp 72-75, 2006.
[15] Artem Khmelinskii, Rodrigo Ventura and Joao Sanches, "A Novel Metric for Bone Marrow Cells Chromosome pairing", IEEE transactions on Biomedical Engineering, Vol 57, No 6, pp 1420 – 1429, June 2010.