Knowl Inf Syst (2017) 52:773–790
DOI 10.1007/s10115-017-1024-6

REGULAR PAPER

An automatic human chromosome metaspread image selection technique

Tanvi Arora[1] · **Renu Dhir**[1]

Received: 2 February 2016 / Revised: 17 January 2017 / Accepted: 28 January 2017 /
Published online: 14 February 2017

Abstract The human chromosome metaspread images are used to generate the karyogram that is used for the diagnosis of the genetic defects. The genetic defects occur due to variation in either the structure of the chromosomes or the number of chromosomes present in the cell. The human chromosome metaspread image selection process is very critical in the karyogram generation task. It is very tedious and time-consuming process and is generally done manually by an expert cytogeneticist. The manual selection results may be biased, and it is possible that the whole search space is not explored to find the best metaspread image. The mood of the cytogeneticist will also greatly affect the selection results. So there is a strong need to automate the process of human chromosome metaspread image selection process. The proposed approach ranks the metaspread images based upon the quality score that is calculated using the count of the chromosomes of various orientations present in the metaspread image. The ranking has been done based upon ordinal ranking process, wherein a unique rank is assigned to each image based upon a set of rules. The rule base aids in the tiebreaking process in case the same quality score is derived for more than one metaspread image. The decision-making process of the expert cytogeneticist has been emulated by using a set of if–then rules. The proposed technique helps to select the best metaspread image, by exploring the complete set of images that can be used for the karyogram generation.

Keywords Chromosomes · Ranking · Feature extraction · Classification · Quality score

1 Introduction

The chromosomes are present inside the nucleus of a cell. They carry the instructions for the synthesis of various proteins. Any alteration in either the structure of the chromosome or the number of chromosomes of an individual can lead to malfunctioning of the proteins,

✉ Tanvi Arora
tanviverma@rediffmail.com

[1] Department of Computer Science and Engineering, Dr. B.R. Ambedkar National Institute of Technology, Jalandhar, Punjab 144001, India

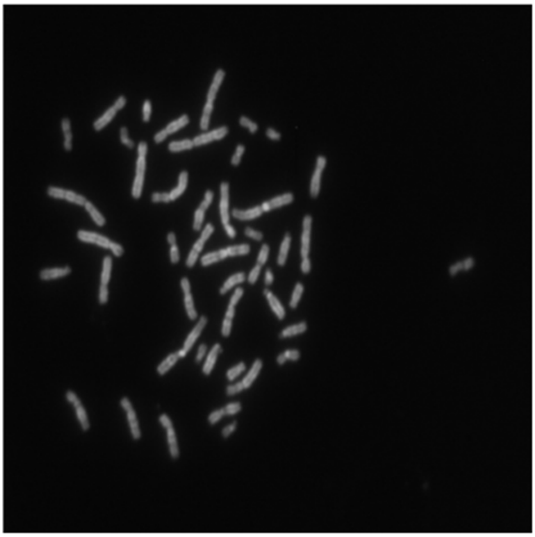


Fig. 1 Human chromosome metaspread image Fig. 2 karyogram generated by an expert cytogeneticist

due to which various diseases or medical conditions may come up that need to be taken care of, and are termed as genetic defects. The genetic defects in an individual are uncovered by studying the number of chromosomes and the structure of chromosomes. A healthy human being is said to have 23 pair of chromosomes [20] out of which 22 pairs are autosomes and the 23rd pair is a sex determining pair of chromosomes that is either XX for females or XY for males. The cytogeneticist studies the structure and number of chromosomes by imaging the cells of the humans during the metaphase of the mitosis phase of cell division, using high-resolution microscopes. The purpose for imaging during the metaphase is that during this stage of cell division the chromosomes are at the longest [2]. The metaspread chromosome images used for the purpose of diagnosis should be of good quality so that the cytogeneticist can retrieve useful and accurate information from them. Figure 1 illustrates the sample image of the human chromosome metaspread captured during metaphase. Figure 2 illustrates the chromosomes that have been arranged by an experienced cytogeneticist in the form of a karyogram, which can be used to uncover the genetic defects.

In order to create a karyogram, the samples of blood, amniotic fluid or maybe tissues of skin are taken. The metaphase chromosomes are extracted from the samples by the process outlined in Fig. 3. The collected sample is first incubated for some days, and during that period, it is treated with phytoagglutinin which leads to increased cell reproduction rate. Towards the end of incubation, the cell division is stopped by treating the sample with colchicines, so as to increase the count of usable metaphase chromosomes. In the second state, the sample is treated with mixture of glacial acid and carbinol so as to fix the cells at a one particular stage of cell division. After this stage, the microscopic slides are taken and the chromosomes are spread for the purpose of observation and imaged using a high-resolution microscope.

The images that are captured for the purpose of karyogram generation contain chromosomes in various orientations, as the chromosomes are non-rigid objects so they may be straight, bent, touching each other or may occur in clusters. The images that contain a large number of touching and overlapping chromosomes are not suitable for karyogram generation. In order to generate the karyograms, the metaspread chromosome images are shortlisted by an experienced cytogeneticist. The images that are shortlisted are selected considering

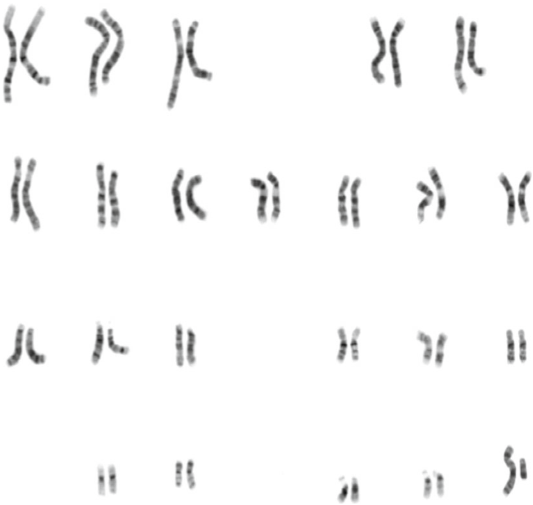


Fig. 2 Karyogram generated by an expert cytogeneticist

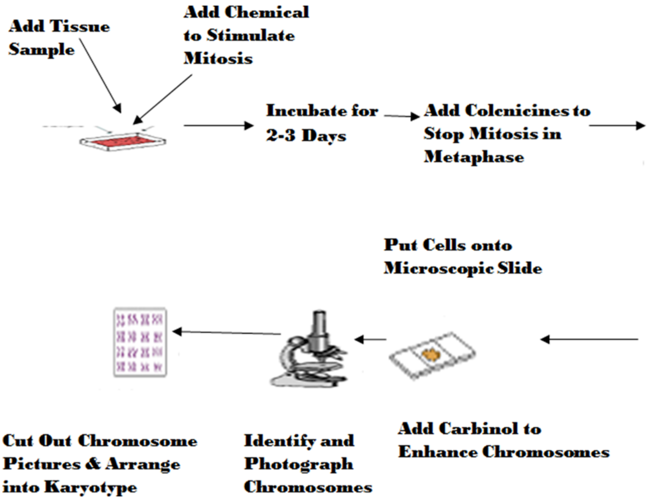

Add Tissue Sample
Add Chemical to Stimulate Mitosis
Incubate for 2-3 Days
Add Colcnicines to Stop Mitosis in Metaphase
Transfer Cells to Tube and Centrifuge to Concentrate in Layers
Put Cells onto Microscopic Slide
Transfer to Tube Containing Glacial Acid
Cut Out Chromosome Pictures & Arrange into Karyotype
Identify and Photograph Chromosomes
Add Carbinol to Enhance Chromosomes

Fig. 3 Metaphase chromosome imaging process

the number of individual chromosomes present in the human chromosome metaspreads as these are the only accurate information providers. In the past many years, much research has been done to disentangle these touching and overlapping chromosomes as reported in [1,3,8,13,14]. The drawback of considering the touching or overlapping or for that matter bent chromosomes for uncovering the abnormalities in chromosomes may lead to inaccurate diagnosis. The inaccuracy comes in as all these models segment or correct the orientation of these overlapping, touching or bent chromosomes based upon hypothesis generation [9,19] and they are not the perfect solutions. In order to get the best karyograms, the human chromosome metaspread images should contain a large number of individual well-separated straight chromosomes.

In order to find human metaspread images having straight chromosomes, they need to be assessed for their quality before proceeding with generation of the karyograms. In the past, most of the work has been done to distinguish between the analysable and non-analysable chromosome images, but these methods are computationally expensive, are slower, are dependent upon human intervention, do not explore the whole search space and are not suitable for processing a large number of images as required for the selection of best human chromosome metaspread images [4,10,12,18,23,24]. MetaSel [21] a metaphase selection tool using Gaussian-based classification is one of the best solution so far proposed.

The rest of the paper is organized as follows: Sect. 2 presents the related work, Sect. 3 presents problem formulation, Sect. 4 contains methodology, Sect. 5 presents results and discussion, and Sect. 6 concludes the paper.

2 Related work

Over the years, few efforts have been made in analysing the quality of the metaspread images before karyogram generation. A method was proposed to classify the selected image as a metaphase chromosome image or a non-metaphase chromosome image by comparing the nine geometric parameters [12]. In this approach, a three-phase detection method is used to find the metaphase chromosome image. In the starting phase, analysis is performed on values of the filtered objects count (FOC) that have gone through band pass. An analysable metaphase chromosome image is possible if it shows sequential high values. In the middle phase, FOC is considered for horizontal distribution. If in this phase it gets consecutive high values, this indicates a possible analysable metaphase chromosome image. In the third phase, a contour following method is applied to completely analyse the image. Thresholding is applied to the image before extraction of the features. The images are classified as metaphase images and non-metaphase images using multivariate classifier. The proposed method is not able to control the quality of the image, is quite slow, does not consider the disoriented chromosomes, and is not able to rank the images so as to prioritize them in order of their quality.

A technique was proposed that counted the number of chromosomes present in the metaspread image [10], the counting was based upon three geometric features, and it had the capability of counting both touching and overlapping chromosomes as well. This system has two phases, namely pre-processing and counting phase. In the pre-processing phase, hysteresis thresholding segments the chromosome objects from the background. Further median filtering method is used to remove salt-and-pepper noise, also to fill the holes of chromosomes and to smooth the chromosome contours so that when thinning is applied extra branches are not created. Thinning operation is performed to obtain the single pixel width skeletons of chromosomes or their clusters. After this, the average width of all the skeletons is calculated. It has been observed that all the chromosomes have consistent width. So all those skeletons that are less than the average width of the skeletons are treated as noise and are not considered. Based upon the same lines, the slight connections are also removed. The procedure is slow and is able to either classify the images as analysable or non-analysable, or rank the metaspread images based upon the count of the chromosomes present.

A method was proposed based upon five features that were a mix of geometry and intensity value-based features. It was able to classify the images into two categorize, viz. analysable images and non-analysable images [24]. It is a five-step process. In the first step, they have taken a digital image and the image quality is enhanced using median filtering. In the second step, the thresholding is applied to remove the high grey values. Third step is region labelling to find connected components and individual pixels are deleted. The fourth step takes the labelled components and computes the features. Then, in the fifth step the computed features are passed to two machine-learning-based classifiers, namely decision trees and artificial neural networks, and the results are further optimized. But it does not consider the chromosomes that are disoriented. Secondly, it is not able to rank the metaspreads in order of quality.

A technique based upon counting connected components is proposed [25], and it is very fast approach and is able to count the disoriented chromosomes as well. In this work, an attempt has been made to extract the count of chromosomes from a metaphase chromosome image that has overlapping chromosomes. For selecting the images, firstly they have used histogram equalization to enhance the image contrast. Then, a binary image is created by using thresholding. In order to remove light and small objects, the binary image has been eroded. But it lacks the capacity to rank the metaspread images or classify them as analysable metaspreads or non-analysable metaspreads.

A method is proposed to classify the metaspread images on the basis of band resolution; the images are classified as either low-band-resolution images which can be used to detect numerical abnormalities or high-band-resolution images that can be used to uncover structural abnormalities [22]. In this work, they have classified the chromosome metaphase images into low and high band resolution considering the shape of chromosomes. In the low-band-resolution images, chromosomes are small in size and are well spread and there is no touching or overlapping, so it is suitable for counting the number of chromosomes. In the case of high band resolution, the chromosomes are long, they may be bent or overlapping, so these chromosomes are used for detecting structural abnormalities. In order to classify the metaphase chromosome images based upon resolution of bands, the metaphase chromosome images are pre-processed based upon grey-level adjustment and Otsu's thresholding, to separate the foreground and background objects. After segmenting the foreground and background objects, the segmented objects are rotated so that they are vertical. The method is not able to rank the images for the purpose of selection.

A software with the name of MetaSel [21] is proposed that ranks the G-banded metaspread images using eight geometric features. It uses the Otsu thresholding [15] method for the purpose of segmentation and classifies the segmented objects into four categories, viz. straight, bent, occluded and noise. It counts the objects of each category and then ranks the metaspread images by considering the count of straight and bent chromosomes. This approach is not suitable for ranking the metaspread images that suffer from intensity inhomogeneity.

Recently, a method for categorizing the metaspread images as analysable and non-analysable is proposed based upon time-delay integration [16]. It is workable in two modes that are online mode and offline mode. It considers five features for the purpose of classification which are based upon geometry and intensity values. They have developed a fully automatic microscope-based image selection system which is based upon the scanning concept of time-delay integration (TDI). They have continuously scanned the image, while the object is moving. The blur is removed by dividing the long exposure time to short exposures. The proposed method can find out more analysable metaphase chromosome images and has more efficiency, and the system can directly provide the high-resolution images to the computer. The technique is not workable for disoriented chromosomes.

The proposed method is compared and contrasted with other state-of-the-art methods developed so far for the purpose of assessing the quality of the metaphase chromosomes images. The comparison has been made on a set of 17 parameters and is presented in Table 1.

Table 1 Comparative performance study with various metaphase selection methods

Feature	R Huber method [12]	Victor Gajendran method [10]	X Wang method [24]	Yan Wenzhang method [25]	Ravi Uttamatanin MetaSel method [21]	Ravi Uttamatanin band classification method [22]	Yuchem Qui's method [16]	Proposed approach
Number of classes	2	0	2	0	4	2	2	5
Control over quality	No	Yes	Yes	Yes	Yes	No	Yes	Yes
Speed	Slow	Slow	Slow	Fast	Fast	Fast	Fast	Fast
Does it rank images	No	No	No	No	Yes	No	No	Yes
Features considered	Geometric	Geometric	Geometric and intensity features	Geometric	Geometric	Geometric and intensity features	Geometric and intensity features	Geometric
Error rate	Not specified	6%	Not given	Not given	10%	15%	Not given	3.5%
Counts number of chromosomes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Counts number of overlapping chromosomes	No	Yes	No	Yes	Yes	No	No	Yes
Number of Features	9	3	5	0	8	8	5	7
Classifier used	Multivariant statistical	Not used	Decision tress and artificial neural networks	Not used	Rule-based Gaussian	Not given	Not given	CFS-CVR

Table 1 continued

Feature	R Huber method [12]	Victor Gajendran method [10]	X Wang method [24]	Yan Wenzhang method [25]	Ravi Uttamatanin MetaSel method [21]	Ravi Uttamatanin band classification method [22]	Yuchem Qui's method [16]	Proposed approach
Workable in case of touching and overlapping chromosomes	No	Yes	No	Yes	Yes	No	No	Yes
Workable when the chromosomes are thin and long	No	Yes	Yes	No	Yes	Yes	Yes	Yes
Segmentation method	–	–	–	–	Otsu method	Otsu method	–	Region-based active contours
Tiebreaking while ranking	–	–	–	–	No	–	–	Yes
Type of images	–	–	–	–	G banded	G banded	–	Any type of banding
Dataset used	–	–	–	–	–	–	–	MFISH ADIR and Q-banded prometaphase

3 Problem formulation

The human chromosome metaspread images are used for the diagnosis of the genetic disorders; therefore, the selection of a very good image having non touching and well-spread chromosome is very important. In order to generate a karyogram from the metaspread image, it should have most of the chromosomes in good orientation, i.e. straight chromosomes. The best features of the chromosomes can only be extracted from straight chromosomes. The straight chromosomes have the clear band patterns, which can be used to detect the structural abnormalities. The images for this purpose are selected manually by human experts, so it is a slow process; the selection of the suitable image depends upon the expertise of the individual; his present state of mind and the whole search space may not be explored. The metaphase image selection is the most time-consuming phase as it is purely dependent on the cytogeneticist.

In the manual image selection process, in order to generate a karyogram, nearly 200 images are analysed by an expert to select the best metaspreads that have clearly separable chromosomes present in it. Since the task is tiring and time-consuming, the experienced cytogeneticist selects the first best 20 metaphase images, so in this way whole of the search space is not explored. The best out of the total lot may never be selected because of manual selection of first 20 metaspread images. Thus, there is a strong need for the development of a system that can emulate the selection criteria of the human expert but by exploring the whole search space.

The automatic system will explore the whole search space, and it will not be dependent upon the availability of the human expert. In the last few years, efforts have been made in developing automatic systems for the selection of metaphase images, but those methods were not practical to implement owing to high cost of processing. Secondly, they were just able to categorize the images into two broad categories as analysable metaspreads and non-analysable metaspread images.

To the best of our knowledge, only one approach for ranking the metaspread images has been found in the literature, in which a software named as MetaSel [21] has been proposed. The software presents a rule-based criteria for ranking the G-banded metaspread images for the purpose of selection of a metaspread image for karyogram generation, and it is a very good effort for ranking the metaspread images. This approach has some limitations. (1) It uses Otsu method for image segmentation, which is not efficient for segmentation of the metaspread images as it contains intensity inhomogeneity [7]. (2) The approach uses Gaussian-based classifier to classify the segmented objects into four classes, viz. straight, bent, overlapping and noise. (3) Only straight and bent chromosomes are used in determining the quality of the chromosomes. (4) The approach cannot efficiently rank the metaspread images that have same number of straight and bent chromosomes.

In the present paper, a model is being proposed that segments the human chromosome metaspread images using the region-based active contour approach as it is capable of efficiently segmenting the images that suffer from intensity in homogeneity. The geometric features of the segmented objects will be extracted, based upon which the objects will be classified into five different classes using correlation-based feature selection and classification via regression classifier. This classifier can classify the objects with very good accuracy. Once the objects have been classified, then based upon the count of the objects of different classes the quality score of the metaspread image will be calculated. Using the quality score, the images will be ranked in the order of most analysable to least analysable and if the quality score results in a tie between two metaspread images, then the tiebreaking is carried out using a set of if–then rules. The proposed approach explores the whole search space and proposes a fully automatic procedure to rank human chromosome metaspread images.

4 Methodology

In order to rank the human chromosome metaspread images, the images undergo the segmentation process using which the objects present in the image are extracted. After the objects are extracted, each object is analysed to extract the geometric features. A set of features are selected from the extracted features using correlation-based feature selection approach. Selected features are used to classify the objects into five classes, viz. straight chromosomes, bent chromosomes, touching chromosomes, overlapping chromosomes and cell residues or noise. Count of each class of objects is calculated and a quality score is computed. Based upon the quality score, the human chromosome metaspread images are ranked. The pseudocode of the proposed approach is depicted below:

```
1. Load the database of images.
2. Segment each image using region-based active contours
3. Calculate the geometric features for each segmented component.
4. Select the correlated features.
5. Classify the segmented objects into five categories, viz. straight chromosome, bent chromosome, touching chromosome, overlapping chromosomes and cell residues or noise for each image using CFS-CVR classifier.
6. Calculate count of each type of objects for each image.
7. Calculate the $Total\,Object\,Count$ using Eq. 2.
8. If $total\ object\ count > 55$ then set Q score = 0 else $Q\ score$ of each image is calculated using Eq. 4.
9. Take all images and $Rank$ them based upon $Q\ score$ using $ordinal\,ranking\,method.$
10. If $rank$ for two or more images is same, then break the tie using the rules given in Table 3.
11. Present the final $ranking$ results.
```

4.1 Segmentation

The metaspread images often suffer from intensity inhomogeneity; therefore, the conventional segmentation approaches like Otsu method and adaptive thresholding are not capable of segmenting them efficiently. In order to perform accurate segmentation of these images, the local intensity values of the nearby regions of the objects are used to find the approximate intensity values along both sides of the contour [5]. The segmentation was carried out using the region-based active contours approach. This approach works very well for the objects that have weak boundaries and intensity inhomogeneity. The segmentation was carried out using MATLAB 2014 software. The proposed approach was able to segment the images quite well; the number of touching chromosomes produced by this segmentation approach is very less.

4.2 Feature extraction

In order to detect the structural and numerical anomalies, the chromosomes present in the metaphase image should be straight. But due to the non-rigid nature of the chromosomes, the chromosomes are present in different orientations, such as straight, bent, touching or overlapping each other in the metaspread image. The metaspread images that have a large number of chromosomes in bent, touching or overlapping orientations are not suitable for detecting the structural anomalies. In order to assess the quality of the metaspread image, the different chromosomes present in it are classified based upon their orientations. In order to classify the chromosomes, 17 geometric features as illustrated in Table 2 have been taken. They have been computed using the MATLAB 2014 software. The features extracted are further normalized so that they have a unit variance and zero mean value, and this process has been carried out using the Waikato environment for knowledge analysis (WEKA tool).

4.3 Feature selection

Out of all the features extracted, some features are independent and few of them are derived features; out of these, some features might not contribute towards the classification results. Therefore, a feature selection approach has been used to search the combination of those features that have the ability to classify the objects into five classes. The selected features will have high discriminating value and will be quite meaningful for the classification purpose. The redundant and irrelevant features have been removed. In this work, the feature selection has been done using correlation-based feature selection (CFS) [11]. It is a simple filter-based algorithm. It uses a heuristic function based upon correlation in finding the relevance of the features. It quickly finds out the redundant, relevant, irrelevant and noisy features. On an average, it may eliminate more than half of the features. In most of the experiments, the classification performance either has been same or has improved by using the reduced feature set as obtained by CFS. There is no requirement of specifying any minimum thresholds or the minimum number of features to be selected; it is a fully automatic algorithm. The importance of the selected features is judged based upon the prediction power of the features and the redundancy associated with them. Those features are chosen that have least inter correlation and more correlation for the class. Following equation illustrates the function that evaluates the subset of features:

$$\text{Merit}_s = \frac{N\overline{p_{\text{cf}}}}{\sqrt{N + N\left(N-1\right)\overline{p_{\text{ff}}}}} \tag{1}$$

where Merit_{s} represents the heuristic-based merit of the subset of N features that have been selected in subset named as s, $\overline{p_{\text{cf}}}$ is the mean value of the feature class correlation and $\overline{p_{\text{ff}}}$ is the average value of the feature to feature inter correlation. The numerator of Eq. 1 highlights that how predictive are the set of features selected and the denominator projects the redundancy amongst the features selected.

WEKA tool has been used for the purpose of feature selection. The CFS attribute evaluator was used with best first searching method that used forward selection heuristic approach and had the stopping criteria after five iterations if no change in subsets takes place. The merit of each subset of features was evaluated using heuristic function using Eq. 1. Here in this study, the subset that has the highest merit of 0.546 was selected. The features of the selected subset are: (1) convex area, (2) minor axis length, (3) solidity, (4) number of branch_pts, (5) number of end_pts, (6) deviation, (7) orientation.

4.4 Classification

The extracted objects from the metaspread images are classified into five classes, viz. (1) straight chromosomes, (2) bent chromosomes, (3) touching chromosomes, (4) overlapping chromosomes and (5) cell residue or noise. The classification of the segmented objects into these five classes has been done considering selected geometric features as described above. Figure 4 shows the chromosomes of these five classes. The straight chromosomes are those

Table 2 Geometric features

Type of features	Feature	Description
Spread	Length	It is the distance between the two extreme end points *Let* (x_1, y_1) *and* (x_n, y_n) *be the two extreme end points the length can be calculated as Length* $= \sqrt{(x_n - x_1)^2 + (y_n - y_1)^2}$
	Area	It is the number of the pixels in the object that have intensity value equal to one $Area = \sum p_i$ *where* p_i *are the pixels of the object having intensity value* = 1
	Convex area	It is the area of the convex hull, where convex hull is the minimum region that is convex and it covers the given region. It is the sum of the pixels in the convex image $Convex area = \sum p_i$ *where* p_i *are the pixels of the convex hull*
	Perimeter	It is the sum of the distance between the adjoining pixels around the boundary of the region $Perimeter = \sum$ *Distance between adjoining pixels* $Distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$
	Equi diameter	It specifies the diameter of the circle with the same area as the region. It is computed as $Equi Diameter = \sqrt{(4 \times Area)/\pi}$
	Major axis length	It is the length of the major axis of the ellipse that has the same normalized second central moments as the region
	Minor axis length	It is the length of the minor axis of the ellipse that has the same normalized second central moments as the region.
	Ratio of minor axis to major axis	It is the ratio of minor axis to major axis. *Ratio of minor axis to major axis* $= \frac{Length\ of\ minor\ axis}{Length\ of\ major\ axis}$
	Solidity	It specifies the pixels in the convex hull that are also in the region it is computed as Area/Convex Area $Solidity = \frac{Area}{Convex\ area}$
	Eccentricity	It specifies the eccentricity of the ellipse that has the same second moments as the region. It is the ratio of the distance between the foci of the ellipse and its major axis length $Eccentricity = \frac{Distance\ between\ foci\ of\ ellipse}{Length\ of\ major\ axis}$
	Extent	Ratio of number of pixels in the region to the number of pixels in the bounding box.$Extent = \frac{Area}{Area\ of\ bounding\ box}$
Shape	Deviation	The pixel values of the medial axis are taken. Then the angle between the adjacent three pixels is calculated as follows: $a(i) = ar\cos\left(\frac{(C_i - C_{i-k}).(C_{i+k} - C_i)}{\Vert C_i - C_{i-k}\Vert.\Vert C_{i+k} - C_i\Vert}\right) sgn \left[\det\left(C_i - C_{i-k} \ldots C_{i+k} - C_i\right)\right]$ if a(i) > 25 then the deviation parameter is value is set to true else false
	Euler number	It is the number of objects in the region minus the number of holes in those objects *Euler number* = *Count of objects in the region-count* of holes in those objects

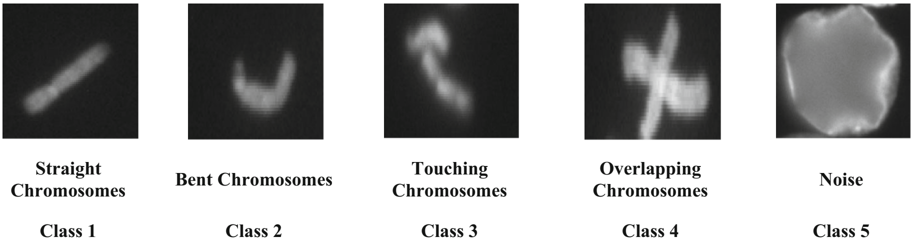
Table 2 continued

Type of features	Feature	Description
	Number of end points	To calculate the number of end points of an image, the image is first skeletonized. Then the number of end points is calculated as follows: $Number\ of\ End\ points = \sum p_i$ *where* p_i *are the pixels obtained by setting the intermediate pixels to 0*
	Circularity	It is the amount of roundness calculated as $Circularity = \frac{4 \times Area \times \pi}{Perimeter^2}$
	Orientation	It is the angle between the x-axis and the major axis of the ellipse that has the same second moments as the region
	Number of branch points	To calculate the number of branch points of an image, the image is first skeletonized. Then the number of branch points is calculated as follows: $No\ of\ Branch\ points = \sum p_i$ *where* p_i *are the pixels that are having 4 connectivity*

objects which are single objects in their desired shape, and the features can be efficiently extracted from them. Bent chromosomes are also single objects, but they may be twisted or not aligned as per their original shape, so it is not possible to accurately extract the features. Touching chromosomes are group of two or more chromosomes that are having their boundary aligned to each other. The disentanglement and then reconstruction of individual chromosomes from touching chromosomes may result in feature distortion. The overlapping chromosomes are cluster of two or more chromosomes that are covering each other. The disentanglement and reconstruction may result in inaccurate or missing information. The cell residues or noise are the objects that are not chromosomes but the residues of cell division or other noise. These objects need to be removed while generating a karyogram. In order to classify the objects extracted into five classes, correlation-based feature selection and classification by regression(CFS-CVR) classifier [6] has been used. The CVR classifier is based upon the model trees; they are a kind of decision trees which have linear regression at the leaf nodes. The model trees are generated by first constructing a simple decision tree; the second stage prunes the tree by replacing the sub-trees by using linear regression. The CVR has been implemented using the random forest algorithm. In this algorithm, a large number of decision tress are built during training time, and in order to classify an object, it is given to each of the trees in the forest. Each tree gives its classification which is treated as a vote for that class; the object is assigned to the class that has the maximum number of votes. This algorithm is a powerful tool for predicting, and it is based upon the law of large numbers so it does not over fit. The accuracy of the algorithm as a classifier depends upon the random inputs and features. It is able to classify the chromosomes into five classes with approximately 95% accuracy.

4.5 Ranking

The ranking of the metaspread images has been done using the ordinal ranking (“ranking,” n.d.). In this, all the objects are assigned a unique number as a rank. The unique rank is assigned to those objects also that have the same value for the ranking parameter. This ranking is done based upon a criteria function to rank the same scoring objects. In the proposed approach, the value of quality score (*Qscore*) is used for the purpose of ranking.



Straight Chromosomes **Bent Chromosomes** **Touching Chromosomes** **Overlapping Chromosomes** **Noise**

Class 1 **Class 2** **Class 3** **Class 4** **Class 5**

Fig. 4 Images of different class of chromosomes

The *Qscore* has been calculated considering the *Total Object Count* of all the objects and the count of objects of different orientations.

The *Total Object Count* is set to sum of count of class 1 and class 2 and twice count of class 3 and class 4. It has been assumed that a minimum number of objects touching and overlapping are 2. The *Qscore* for the images for which the *Total Object Count* exceed 55 is set to zero. So the value greater than 55 means that these images have lot of noise and have not been segmented properly.

$$\textit{Total object count} = \textit{class 1} + \textit{class 2} + 2 * \textit{class 3} + 2 * \textit{class 4} \tag{2}$$

For the images that have the *Total Object Count* less than 55, the *Qscore* has been calculated considering the number of chromosomes of each of the four classes; different weights are assigned to each of the attributes that are used to measure the quality score. The different weights are assigned considering the accuracy of features that can be extracted from the segmented chromosomes, more the accuracy of the extracted features more the weight assigned. The class 1 objects are assigned a weight of 0.60 as they are already in the best possible orientation and the features can be extracted efficiently. The class 2 objects are assigned a weight of 0.20 as effort is needed to bring them to the desired orientation and the accurate feature extraction may not be possible. The class 3 objects are assigned a weight of 0.10 and class 4 objects are assigned a weight of 0.05 as considerable amount of effort is needed to correct their orientation, and the reconstruction of individual chromosomes may not result in accurate feature extraction. The *Qscore* is calculated by summation of the normalized count values of each class by multiplying them by their weights.

$$Qscore = f\,(SC, BC, TC, OC) \tag{3}$$

$$Qscore = \sum_{i=1}^{4} w_i \left[\left(x_i - \overline{x_i} \right) / s_i \right] \tag{4}$$

where x_i is count of the chromosomes of class, $\overline{x_i}$ is mean value of x_i, s_i is standard deviation of x_i and w_i is the weight assigned to that class.

The rank calculation based upon the quality score is done using the algorithm as described in “Methodology”. The ranking approach aims at placing the most analysable human chromosome metaphase image at the first position based upon the quality score parameter and the least analysable image at the bottom, so that the cytogeneticist can choose the most analysable metaspread image with least effort. The main effort that the algorithm puts in is in resolving the tie amongst different metaspread images based upon the quality score. The automatic tiebreaking is strictly done as per the tiebreaking done by the cytogeneticist while manual selection.

Table 3 Tiebreaking rules for ordinal ranking

Rule no	Situation	Condition	Action
1.	If the quality score is same for two or more images	Check the count of straight chromosomes	The image having more count of straight chromosomes gets a better rank
2.	If count of straight chromosomes is same	Check count of bent chromosomes	The image having less count of bent chromosomes gets a better rank
3.	If count of straight and bent chromosomes is same	Check for count of touching chromosomes	The image having less count of touching chromosomes gets a better rank
4.	If count of straight, bent and touching chromosomes is same	Check for count of overlapping chromosomes	The image having less count of overlapping chromosomes gets a better rank
5.	If the count of all types of chromosomes is same	Check count of noisy objects	The image having less count of noisy objects gets a better rank
6.	If count of chromosomes and noisy objects is also same	Check the order in which the images appear	The image that appears ahead gets a better rank

In case the quality score for more than one image comes out to be same, then the image that has more number of class 1 objects is assigned a higher rank. But in case the number of class 1 objects is also equal, then count of the class 2 objects is considered. In this case, the image that has less number of class 2 objects is assigned a higher rank. Still if the class 2 objects are equal, then same procedure is adopted for class 3 objects and class 4 objects also. Meaning in any case the image that has more number of class 1 objects and less number of class 2, class 3, class 4 and class 5 objects are to be assigned a higher rank. But still if the number of objects of all the five classes is same, then the image that appears first in the list is given a higher rank (Table 3).

5 Results and discussion

The 200 DAPI images of the ADIR dataset were used for the purpose of ranking. The results of the ranking were compared with those of the ground truth. The 17 metaspread images that were marked as difficult to analyse in the ground truth database occupied the least ranked positions by the proposed method as well except three images which occupied 34th, 92th and 96th position. Apart from this, four other images had a mismatch in the rank; out of 200 images, 193 were ranked correctly. So the proposed model was able to correctly rank the images with the accuracy of 96.5%. The results obtained with the proposed method have been compared with the results obtained by applying the technique proposed by the MetaSel for ranking [21]. The top five metaspreads of our approach completely matched the top five metaspreads as ranked by the experienced cytogeneticist, whereas MetaSel could only rank 3 metaspread images correctly out of 5. The performance comparison of proposed approach with MetaSel approach is presented in Table 4.

MetaSel method does not consider the total count of chromosomes while calculating the rank. The metaspreads did not match correctly with the proposed method, and with the ranks

Table 4 Table performance comparison

Image	Number of class 1 objects	Number of class 2 objects	Number of class 3 objects	Number of class 4 objects	Rank by MetaSel	Rank by proposed approach	Rank by expert
1	24	3	5	1	1	1	1
2	22	12	8	1	2	4	4
3	18	4	2	8	3	3	3
4	18	3	6	8	4	5	5
5	17	7	7	2	5	2	2

Table 5 Comparison with MetaSel

Parameter	MetaSel	Proposed approach
Segmentation technique used	Otsu thresholding	Region-based active contours
Classifier used	Rule-based Gaussian classifier	CFS-CVR classifier
Ranking	Semi-automatic, was not able to handle ties	Fully automatic ordinal ranking
Number of classes of objects segmented	4	5
Ranking based upon	Class 1 and class 2 chromosomes	Quality score computation based upon weighted consideration for all four classes of chromosomes
Applicable to	G-banded chromosomes	Any type of metaspread images
Features considered for classification	Eight	Seven
Accuracy	93.19%	96.5%

of experienced cytogeneticist, they had a large number of objects as chromosomes. The proposed approach was not able to give the same ranks as it assigned a quality score of 0 to those images which had a large number of objects as chromosomes.

The proposed method is fully automatic as it does the ranking based upon the rules. The tiebreaking process is also very robust. The mistakes done during the segmentation procedure are also taken care of by considering the total number of objects of classes 1–4 while calculating the rank. Table 5 shows the comparison of the proposed approach with the MetaSel approach. The proposed approach outperforms the MetaSel method on several parameters.

The accuracy of the proposed approach is 96.5%, the best possible segmentation, and classification approach has been applied to segment and classify the chromosomes. An accuracy of 100% can be achieved if a robust reconstruction approach is formulated to reconstruct the bent, touching and overlapping chromosomes.

6 Conclusion

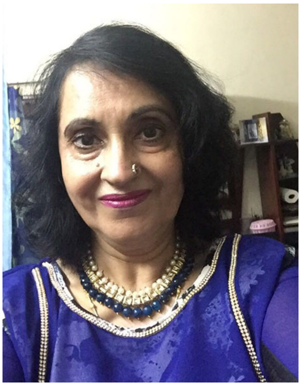
In this work, an effort has been made to rank the human chromosome metaspread images, so that the work of the cytogeneticist is reduced. The objects present in the metaspread are categorized into five different classes. Based upon the count of the different number of chromosomes of each class, the metaspread images are assigned a rank from most analysable to least analysable images based upon a quality score. The ranking has been done using the ordinal ranking process. In case the quality score comes out to be same, then based upon rules the tiebreaking has been done. This automated approach has contributed in the following ways: (1) the proposed method can handle any type of images as it is based upon geometric features. (2) Since the approach is automatic, the whole set of images will be examined to bring the best image out. (3) The image selection time will be reduced. (4) The unanalysable images will rank lower and time will not be spent on analysing those images. (5) The processing time for karyogram generation will be reduced. (6) The experimental results show that the proposed approach is able to give 96.5% accurate results. In future, the proposed technique can be integrated in the automatic karyogram generation process to speed up the process.

References

1. Agam G, Member S, Dinstein H (1997) Geometric separation of partially overlapping nonrigid objects applied to automatic chromosome classification. IEEE Trans Pattern Anal Mach Intell 19(11):1212–1222
2. Alberts VV, Hillie KT, Demanet CM (2000) Atomic force microscopy imaging of polycrystalline CuInSe2 thin films. J Microsc. 197(Pt 2):206–215. doi:10.1046/j.1365-2818.2000.00652.x
3. Arora T, Dhir R (2014) An efficient segmentation method for overlapping chromosome images. Int J Comput Appl 95(1):29–32. doi:10.5120/16560-4861
4. Arora T, Dhir R (2015) A review of metaphase chromosome image selection techniques for automatic karyotype generation. Med Biol Eng Comput. doi:10.1007/s11517-015-1419-z
5. Arora T, Dhir R (2016a) A novel approach for segmentation of human metaphase chromosome images using region based active contours. Int Arab J Inf Technol, pp 1–5
6. Arora T, Dhir R (2016b) Correlation based feature selection & classification via regression of segmented chromosomes using geometric features. Med Biol Eng Comput. doi:10.1007/s11517-016-1553-2
7. Arora T, Dhir R (2016c) Segmentation approaches for human metaspread chromosome images using level set methods. In: International conference on mass data analysis of images and signals MDA 2016 in New York
8. Choi H, Bovik AC, Castleman KR (2006) Maximum-likelihood decomposition of overlapping and touching M-FISH chromosomes using geometry, size and color information. In: Twenty-eighth annual international conference of the IEEE engineering in medicine and society, New York
9. Devaraj S, Vijaykumar VR, Soundrarajan GR (2013) Leaf biometrics based karyotyping of G-band chromosomes. Int J Hum Genet 13(3):131–138
10. Gujendran V, Rodriguez JJ (2004) Chromosome counting via digital image analysis. In: Proceedings of international conference on image processing. IEEE, pp 2929–2932
11. Hall M (1999) Correlation-based feature selection for machine learning. PhD Thesis., Department of Computer Science, Waikato University, New Zealand (April)
12. Huber R, Kulka U, Lörch T, Braselmann H, Bauchinger M (1995) Automated metaphase finding: an assessment of the efficiency of the METAFER2 system in a routine mutagenicity assay. ELSEVIER Mutat Res 334:97–102
13. Karvelis P, Likas A, Fotiadis DI (2010) Identifying touching and overlapping chromosomes using the watershed transform and gradient paths. Pattern Recogn Lett 31:2474–2488. doi:10.1016/j.patrec.2010.08.002
14. Munot MV (2011) Automated Karyotyping of Metaphase Cells with Touching Chromosomes. Int J Comput Appl 29(12):14–20
15. Otsu N (1979) A threshold selection method from gray-level histograms. In: IEEE transactions on systems, man, and cybernetics, pp 62–66. doi:10.1109/TSMC.1979.4310076
16. Qiu Y, Chen X, Li Y, Chen WR, Zheng B, Li S, Liu H (2013) Evaluations of auto-focusing methods under a microscopic imaging modality for metaphase chromosome image analysis. Anal Cell Pathol 36:37–44. doi:10.3233/ACP-130077
17. ranking. (n.d.). Retrieved 24 Jan 2016. http://www.merriam-webster.com/dictionary/ranking
18. Shippey G, Carothers AD, Gordon J (1986) Operation and performance of an automatic metaphase finder based on the MRC fast interval processor. J Histochem Cytochem 34(10):1245–1252
19. Somasundaram D, Palaniswami S, Vijayabhasker R, Venkatesakumar V (2014) G-band chromosome segmentation, overlapped chromosome separation and visible band calculation. Int J Hum Genet 14(2):73–81
20. Tjio JH, Levan A (1956) The chromosome number of man. Genetics 10(6):80–85
21. Uttamatanin R, Yuvapoositanon P, Intarapanich A, Kaewkamnerd S, Phuksaritanon R, Assawamakin A, Tongsima S (2013) MetaSel: a metaphase selection tool using a Gaussian-based classification technique. BMC Bioinform 14:S13. doi:10.1186/1471-2105-14-S16-S13
22. Uttamatanin R, Yuvapoositanon P, Intarapanich A, Kaewkamnerd S, Tongsima S (2013) Band classification based on chromosome shapes. In: 13th international symposium on communications and information technologies (ISCIT), pp 464–468
23. Van Den Berg HTCM, De France HF, Habbema JDF, Raatgever JW (1981) Automated selection of metaphase cells by quality. Cytometry 1(6):363–368. doi:10.1002/cyto.990010602

24. Wang X, Li S, Liu H, Wood M, Chen WR, Zheng B (2008) Automated identification of analyzable metaphase chromosomes depicted on microscopic digital images. J Biomed Inform 41(2):264–271. doi:10.1016/j.jbi.2007.06.008
25. Wenzhong Y (2009) A counting algorithm for overlapped chromosomes. In: 3 rd international conference on bioinformatics and biomedical engineering. IEEE, pp 1–3

Tanvi Arora received her B.Tech degree in Computer Science and Engineering from Punjab Technical University, Jalandhar, Punjab, India, in 2002, M.Tech degree in Information Technology from Punjab Technical University, Jalandhar, Punjab, India, in the year 2007. She is currently pursuing Ph.D. in Computer Science and Engineering from Dr. B.R. Ambedkar National Institute of Technology, Jalandhar, Punjab, India, and working as Associate Professor in the Department of Computer Science and Engineering at Baddi University of Emerging Sciences and Technology, Baddi, Himachal Pradesh, India. Her teaching and research interests include image processing, pattern recognition, machine learning, data mining and network security.



Dr. Renu Dhir received her B.Tech degree in Electrical Engineering from Guru Nanak Dev Engineering College, Ludhiana, Punjab, India, in 1983, M.Tech degree in Computer Science and Engineering from TIET Patiala, Punjab, India, in the year 1997, and Ph.D. degree in Computer Science and Engineering from Punjabi University, Patiala, Punjab, India, in 2007. She is currently working as Associate Professor at Dr. B.R. Ambedkar National Institute of Technology, Jalandhar, Punjab, India. Her teaching and research interests include image processing, pattern recognition, machine learning and network security.