Automatic karyotyping based on computational vision and intelligent classifiers

Pedro F. R. Ferraz

Instituto Superior Técnico
Lisboa,Portugal,2010 (e-mail: FerrazPedro@live.com.pt).

Abstract: The objective of this work is to create an intelligent system to support the construction of the karyotype through the automatic classification of chromosomes. The karyotype construction, or karyotyping, is a part of a type of analysis realized in laboratory at genetic level. This process part, is time consuming and requires highly specialized personal in live science. Thus, this work uses techniques and algorithms of image processing and intelligent systems of decision (artificial neural networks and fuzzy logic) to implement a method that can build automatically the karyotype. Image processing techniques were developed to characterize some of the descriptors, particularly in terms of obtaining the middle axis of the chromosome. Features were tested, getting results without resorting to the location of the centromere. The performance of the two classification systems was tested and compared.

Keywords: Karyotyping, Classification, Artificial Neural Networks, Fuzzy Logic, Clustering, Image Processing.

1. INTRODUCTION

1.1 Chromosomes

Chromosomes are tiny structures found in the nucleus of eukaryotic cells that containing a vast amount of genetic information (Deoxiribonucleic Acid) about each individual. These structures are usually, in the human case, consisted of 22 asexual pairs plus a XX pair for the female individuals, or a XY pair for male ones. One should never discard the possibility of being more or less than 2 chromosomes per pair, since there are diseases that cause these unusual cases, such as the trissomy 21 (3 chromosomes present in pair No. 21).

On the structural level, chromosomes are composed of a filamentous DNA spiral whose density along the body of the chromosome may vary, since there is a specifically more concentrated area labeled by centromere. In the human case, each chromosome has one centromere, which location is one of the features that identify the type of chromosome. Regarding the classification of chromosomes by their centromere, they are distinguished as follows:

• metacentric – have the centromere in the middle, dividing the body of the chromosome into two almost identical parts;

• submetacentric – have the centromere away from the center, dividing the chromosome into two parts with different sizes;

• acrocentric – have the centromere very close to the extremity

the sections of the chromosome that do not constitute the centromere are identified as q-arm in the case of the section of greatest length and p-arm in the case of smaller length. Note that in the case of the metacentric chromosome that fact is not so evident because this has the p-arm with almost the same length than the q-arm.

1.2 Karyotyping

Karyotyping is part of a specific type of genetic analysis performed in the laboratory that consists of the ordering of chromosomes in pairs obtained from a starting image where they appear naturally disordered.

Karyotyping begins with obtaining multiple images under a microscope, of the chromosomes spread along the blade, which main objective is the organization of the chromosomes according to a standardized way that in the future can be adequately analyzed structural defects or abnormal numbers of chromosomes that will lead to different conclusions on biological level. The organization of the chromosomes is a purely digital process, this is, the actual chromosomes present in the blade are not moved at time of karyotyping, only processing the obtained image (s). The karyotype should be organized into groups (Denver Groups) as shown in Table 1 in the positions seen in Figure 1 [1]. In this figure is also illustrated an example of the image initially obtained under the microscope Fig. 1.a and the image of the processed karyotype Fig. 1.b, as well as the G-bands present in the chromosomes. For classification by pairs (within groups), resorts to the banding profile, including locations and intensities of the G-bands.

Table 1 - Properties for classification of chromosomes into groups

Group	Pair	Size	Centromere position
A	1,2,3	Biggest	Metacentric
B	4,5	Big	Submetacentric
C	6,7,8,9,10,11,12	Median	Submetacentric
D	13,14,15	Median	Acrocentric
E	16,17,18	Small	Submetacentric
F	19,20	Small	Metacentric
G	21,22	Smallest	Acrocentric
X	Same properties than group C		
Y	Same properties than group G		

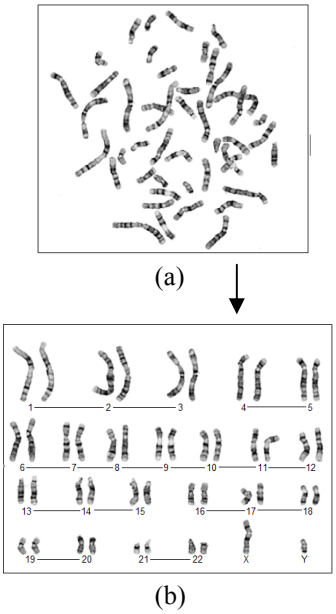


Fig. 1 – (a) the microscope image obtained; (b) image of the karyotype.

2. IMPLEMENTED ALGORITHM

The algorithm implemented in this work consists of two main parts, image processing and classification. In the first part, first the original image is binarized, shown in grayscale (256 different levels), followed by a segmentation process, filtration and subsequent extraction of descriptors. In the classification, independent rating systems are resorted to classify the chromosomes. This algorithm is shown schematically in the appendix of this report.

The algorithm implemented in this thesis considers some assumptions: Images from already built karyotypes were analyzed with low noise level, taking into account that the chromosomes are in the upright position. Regarding the resolution, these images have 1416 by 1040 pixels, this is, about 1.47 mega pixels where one chromosome has between 30 to 300 pixels wide.

3. IMAGE PROCESSING

The binarization of the images was carried out with an imposed threshold of 0.95, assuming that all of the 255 levels in an image are distributed between 0 and 1. The assigned value to the threshold is an value obtained empirically showing great resilience in the database used here, as these pictures show a very clear background with a lot of pixels at its highest level. There was tempted to use the Otsu method [1], however, given the properties of these images and that it is based on the histogram, this method failed to automatically obtain the level to be imposed as a threshold. Otsu's method is best used for an image which background has a more significant variance in relation to the distribution of gray levels, which is not the case.

In Fig. 2 is possible to compare two types of threshold: automatic Otsu level and manual imposed value of 0,95.

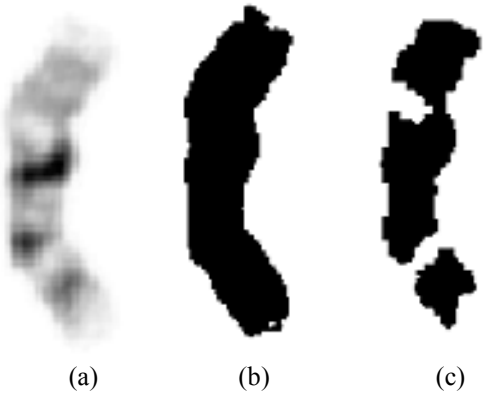


Fig. 2 – (a) Original chromosome image; (b) Chromosome image with imposed threshold of 0.95; (c) Chromosome image with Otsu threshold

After binarization of the image of the karyotype, it will be necessary to recognize and locate the different chromosomes present in the image while removing objects candidates to chromosomes but which are not. This operation corresponds to the segmentation of the binary image, having been, to this end, developed the "segmentation" function.

Analyzing the segmented chromosomes, it appears that they still may have some gaps arising from the binarization process. This problem occurs because some bands present in the chromosome are very right and thus are below the imposed threshold. In order to circumvent this problem, an filling algorithm is applied, where all the interior gaps left in the binarization are filled. In order to be able to eliminate the peripheral notches caused by the same situation described above, and given that the filling algorithm eliminates only gaps inside a closed region, a median filter that smooths the edges of chromosomes is applied. Fig. 3 shows the effect caused by filling algorithm and median filter.

(a) (b) (c) (d)

Fig. 3 – (a) Original chromosome image; (b) chromosome image with imposed threshold of 0,95; (c) after applied filling algorithm; (d) after applied median filter.

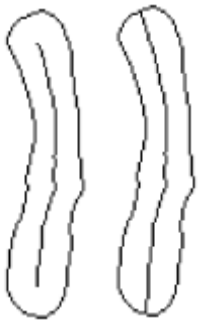
3.1 Feature Extraction

The area considered here is the one which corresponds, in a binary image, to the amount of pixels present in a given region, this is, in this case, it will be the amount of pixels in the region that defines the chromosome. The perimeter of a given region of a binary image is defined as the number of pixels that lie on the border of that region. To get the image that define perimeter, was used an function of *MATLAB* called *remove* that basically left the contour of an binary image. An example of two images where the area and perimeter are extracted is presented in Fig. 4, where the area have a value of 3338 pixels and perimeter 286 pixels.

(a) (b)

Fig. 4 – (a) pre-processed binary image; (b) same image after applied remove function

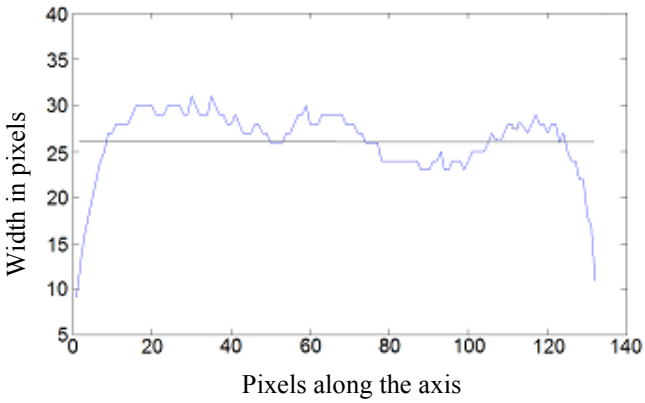
The middle axis is the line that defines the center of the chromosome along its entire length. To obtain the middle axis, the binary image that defines the chromosome was used and applied a function of erosion created specifically for this purpose. It was also necessary to develop a function to cut the branches left by the previous algorithm and another that ensures continuity of the line that defines the middle axis. Finally, it was necessary to extend the line until the chromosome ends, applying an extrapolation algorithm. In Fig. 5 is visible an example of middle axis algorithm applied.



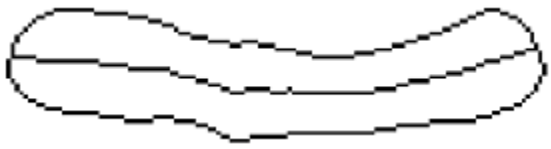
(a) (b)

Fig. 5 – (a) binary image of chromosome after applied erosion function and respective perimeter; (b) image representing middle axis extended to all chromosome body and respective perimeter.

In order to characterize the chromosome as to its width, it was necessary to use a methodology to measure the width in pixels along its entire length. To this end, the middle axis and perimeter previously obtained were resorted, by which orthogonal line segments to the axis can be measured in pixels, the distance between the lines that define the perimeter (see Fig.6). Using the lines orthogonal to the middle axis, it was possible to obtain a profile of banding, obtaining the average gray level of the body of the chromosome for each middle axis (see Fig.7).

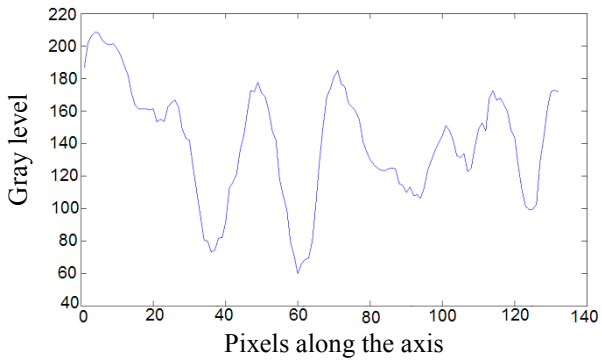


(a)

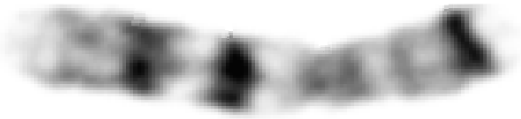


(b)

Fig. 6 – (a) function that represents the width profile of chromosome perimeter and middle axis presented in (b).



(a)



(b)

Fig. 7- (a) function that represents the banding profile of chromosome presented in (b).

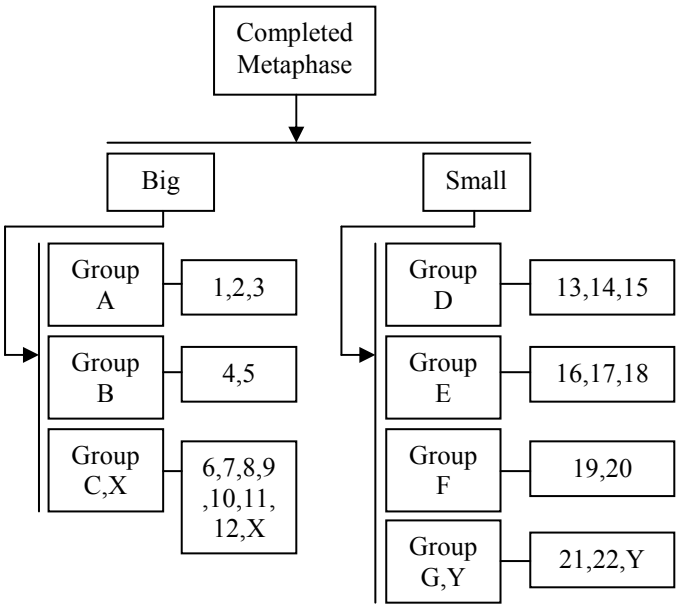


Fig. 8 - Scheme used for classification by stages

4. CLASSIFICATION

4.1 Classification methodology used

Two distinct types of classification were used: fuzzy logic algorithm by optimizing the Gustafson Kessel Clustering and Artificial Neural Networks.

Similar to Wang et al. [2], various independent rating systems linked together were used to provide a method of tree classification. This classification system starts from a classification by size, where only one can distinguish the large type chromosomes of the small type chromosomes, whereas in the group of large chromosomes are part of it subgroups A, B, C and X and in the group of small chromosomes, the subgroups D, E, F, G and Y. Using two classifiers independently, a classification of subgroups was obtained (A to G, X and Y), being a classifier for chromosomes that comprise the set of large and another classifier for chromosomes contained in the small group , and at this stage the subgroup X is integrated in subgroup C and subgroup Y in G, because of the great similarity between the chromosomes of these subgroups.

After these first two stages of classification, a separate listing is used for each subgroup in order to distinguish the pairs belonging to each of these subgroups, as illustrated in Fig. 8.

4.2 Image features used

To classify by size, a set of descriptors solely related to the geometry presented by the chromosome was used. Fig. 9 represents inputs and outputs used in the classification system.

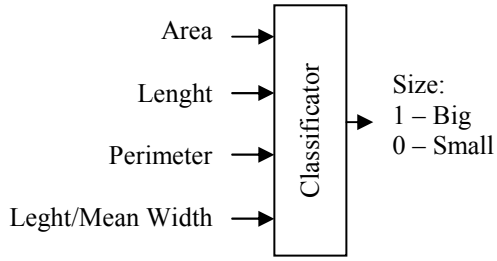


Fig. 9 - Schematic of the inputs and outputs used in the classification system for the step of classification by Size.

All features used here, were first normalized between zero and one for each metaphase using as reference the maximum value for each feature in each metaphase.

In the classification by groups and pairs, the features considered were related not only to the structure of chromosomes, but also with the banding profile.

An histogram of gray levels present in the original image of the chromosome was extracted, that on its turn divided 256 levels into just 4. The information was used on the histogram using four inputs in neural network, where each of the four symbolizes the number of pixels contained in the respective grade level.

In group classification are used 15 average gray levels, from 15 regions defined as being equally spaced along the middle axis. While still respecting the descriptors related to the banding, the location of the centroid of the banding

profile was calculated, this is, indicates the location of the point where the area function that represents the banding profile is equal both to is left as on is right (see Fig. 10.c).

With regard to the outputs of the classifiers, these are identified by number from 1 to 3 for the case of the classification system for groups of chromosomes previously classified as large, 4 to 7 for groups of chromosomes previously classified as being small, as illustrated by Fig. 11. The inputs and outputs used in this classifier are shown in Fig. 11.

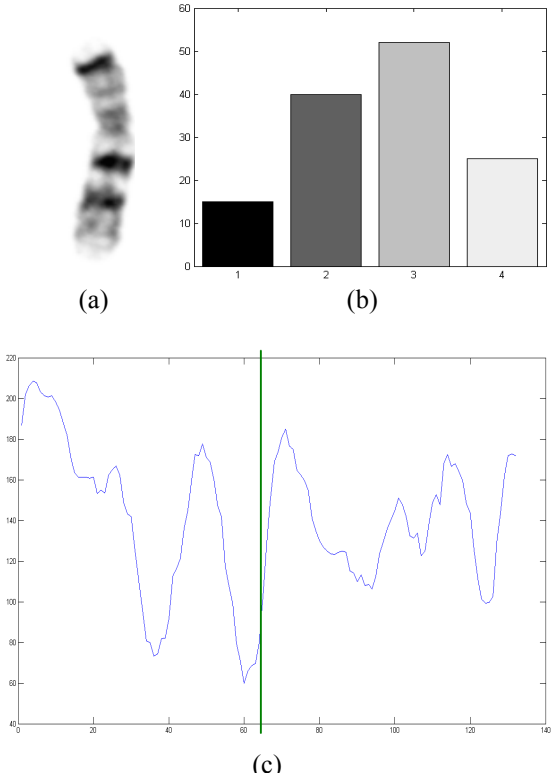


Fig. 10 – (a) Original chromosome image; (b) Histogram with the 4 used values; (c) Centroid function location that represents the bandig profile

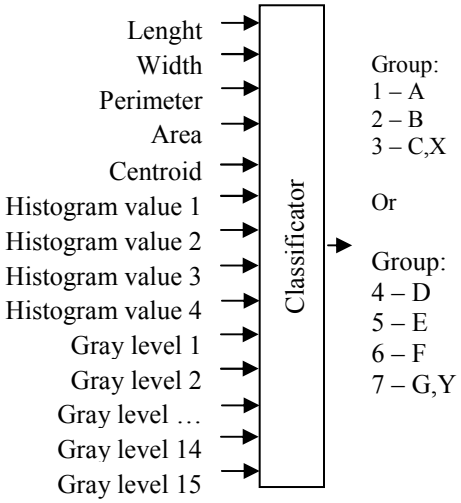


Fig. 11 - Schematic of the inputs and outputs used in the classification system to the step of classification by Group.

5. RESULTS

In general, analyzing the ANN compared to systems that use the FL, it was possible to conclude that systems that use FL always have better results. This may be due to the clustering technique working best for the descriptors used here for classification. In total, one can assume that the results are satisfactory, given that was achieved, a good level of classification for the majority of cases always using the non-complex classification systems (except the rating by pars of group C, X). Table 2 presents a summary of the results obtained for both classification systems and that justified these conclusions.

Table 2 - Relative errors for each system of classification obtained with ANN and Fuzzy Logic.

	Relative classification error of test data (%)	
	RNA	LF
Size	1,2	0,9
Groups of big chromosomes	4,9	3,1
Groups of smal chromosomes	13,7	7,6
Pairs of group A	2,6	0,0
Pairs of group B	0,0	0,0
Pairs of group C,X	28,0	20,0
Pairs of group D	5,1	2,6
Pairs of group E	5,1	5,1
Pairs of group F	0,0	0,0
Pairs of group G,Y	0,0	0,0

6. CONCLUSIONS

The results obtained in this work are satisfactory, regarding the comparison made with a reference article [2].

The vast majority of authors use the location of the centromere to classify chromosomes, since this is an indicator commonly used to identify the type of chromosome. Using the knowledge of experts in the area, it is known that in practice they can identify the type of chromosome only by their banding, so possibly the location of the centromere should not be a crucial fact in the classification of chromosomes. The location of the centromere is not easy to achieve both computational ([3], [4]) and human level. With the results obtained in this work, since the location of the centromere wasn't used, the idea that the centromere is not a crucial feature in the classification of chromosomes can be reinforced.

In the classification by size, good results were achieved thanks to the inputs used in the classification system, since there is an almost direct relationship between

the size of the chromosome and its length (one of the inputs considered).

The results of classification by pairs of groups B, F and G, Y, may be explained by the significant difference that exists between the chromosomes of these groups and because of the classification systems used to identify these cases have two or three types of chromosomes only. In the case corresponding to the pair of chromosome 4, it appears with a darker tone compared to the chromosome pair belonging to No 5. Chromosome No. 19 differs by its profile banding, with a light color and homogeneous, except for a band clearly visible in the center, whereas the chromosome No. 20 has a larger amount of dark bands. Chromosomes 21 and No. 22 are distinguished by the location of the darkest band and differentiate very well from the Y chromosome, since this does not present a band with such a high contrast compared with the others.

The classification of group C, X, shows clearly a worse outcome. This group, unlike the groups that have a high rating, features strikingly similar chromosomes, so it is natural that there is more difficulty in distinguishing the different pairs. In the article of reference [2], this was also the group with worse outcomes. This poor result compared to the other may possibly also be explained by the large variety of chromosomes to be classified, as in this case, for the same rating system, qualify eight chromosomes corresponding to different pairs, and not just two or three as in cases of higher success.

7. FUTURE WORK

For future work, to complement the work in this thesis, new developments both in terms of image processing and in classification are suggested.

It´ s very frequent that in an original image obtained at the microscope, the chromosomes are overlapped, so that this problem has challenged some authors to find a way to not only digitally separate chromosomes that are overlapped, but also those who are just touching side by side or top by top.

One of the assumptions imposed on the classification of chromosomes in this work is that the chromosomes are already in its correct orientation. There is suggested a way to align the chromosomes, producing a first classification so as to guide them with regard to their polarity.

The problems in obtaining a good result of classification for group C, X, suggest that this group has to be analyzed in a different way, possibly looking for new descriptors that can be applied here or even applying an intermediate stage of classification, dividing into smaller subgroups.

The application of new methods of classification is suggested, which eventually might be created and applied to this type of problem.

8. REFERENCES

[1] N. Otsu, "A threshold selection method from gray-level histograms," *Systems, Man and Cybernetics*, vol. 9, pp. 62-63, 1979.

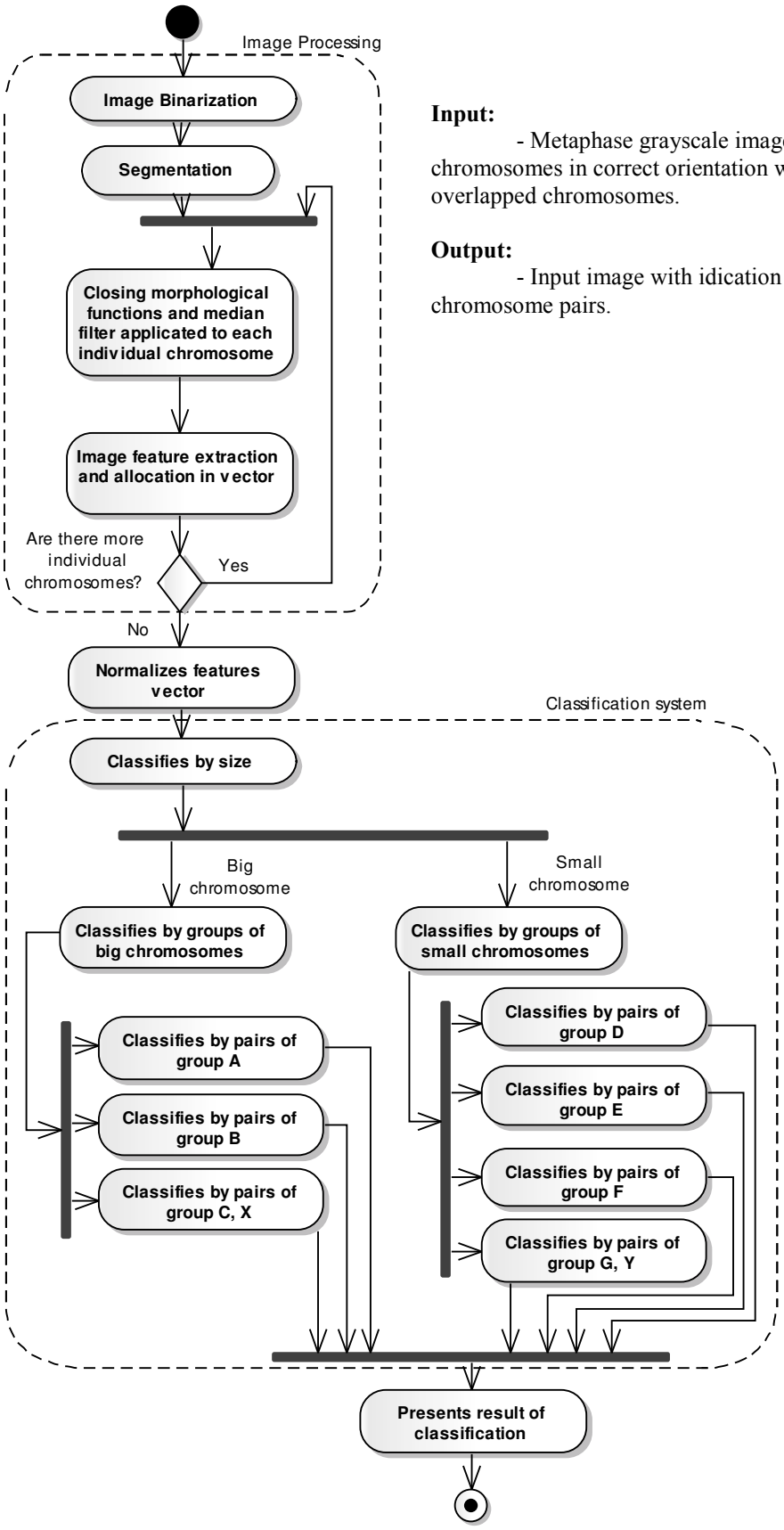
[2] X. Wang, et al., "Automated classification of metaphase chromosomes: Optimization of an adaptive computerized scheme," *Journal of Biomedical Informatics*, vol. 42, pp. 22-31, 2009.

[3] J. M. Cho, "Chromosome Classification Using Backpropagation Neural Networks," *Engineering in Medicine and Biology Magazine*, vol. 19, pp. 28-33, Jan. 2000.

[4] B. Legrand, C. S. Chang, S. H. Ong, S.-Y. Neo, and N. Palanisamy, "Chromosome classification using dynamic time warping," *Pattern Recognition Letters*, vol. 29, pp. 215-222, 2008.

[5] A. M. Badawi, K. G. Hasan, E.-E. A. Aly, and R. A. Messiha, "Chromosomes classification based on neural networks, Fuzzy rule based, and template matching classifiers," *Micro-NanoMechatronics and Human Science*, vol. 1, pp. 383-387, 2003.

9. APPENDIX



Input:

- Metaphase grayscale image of chromosomes in correct orientation without overlapped chromosomes.

Output:

- Input image with idication of the chromosome pairs.

Fig. 12 - Diagram of the implemented algorithm (UML 2.0)