COMPUTATIONAL PROBLEMS OF ELECTRICAL ENGINEERING

Vol. 12, No. 2, 2022

EXTRACTION OF IDEOGRAM FEATURES FOR DIAGNOSING CHROMOSOMAL ABNORMALITIES

Oleksii Pysarchuk, Yurii Mironov

Department of Software Enineering, National Aviation University, Kyiv, Ukraine

PlatinumPA2212@gmail.com, yuriymironov96@gmail.com

https://doi.org/10.23939/jcpee2022.

Abstract: This paper proposes an approach to the detection and extraction of specific features in an ideogram image. Ideogram is a depiction of a healthy chromosome [1] used in a karyotyping process - a procedure designed to diagnose chromosomal abnormalities [2].

Extraction of ideogram features is a part of a general algorithm for the detection of chromosomal abnormlities [3, 4]. According to the general algorithm, both chromosomes and ideograms have to be parsed and converted into a single data format for further comparison.

The image of the ideogram is the input data for the algorithm of the extraction of ideogram features, which is proposed in this paper. The output is a data structure containing ideogram properties. A software prototype has been developed to verify the algorithm efficiency.

Key words: Computer Vision, Feature Extraction, Image Processing, Noise Removal.

1. Introduction

Chromosomal disorders, such as Down syndrome, can cause multi-domain disability and are correlated with more frequent deaths before the age of five [5]. Moreover, chromosomal disorders are associated with a higher risk of miscarriage [6].

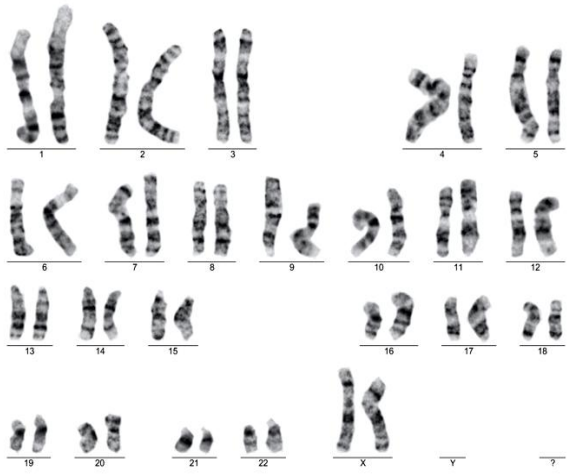

Fig. 1. Human karyogram.

In order to diagnose chromosomal disorders, a procedure called karyotyping is conducted. Karyotyping is a process creating a karyogram, that is, of identifying and arranging human chromosomes (Fig. 1).

During the karyogram arrangement, each chromosome is compared to the ideal chromosome image called an ideogram (Fig. 2).

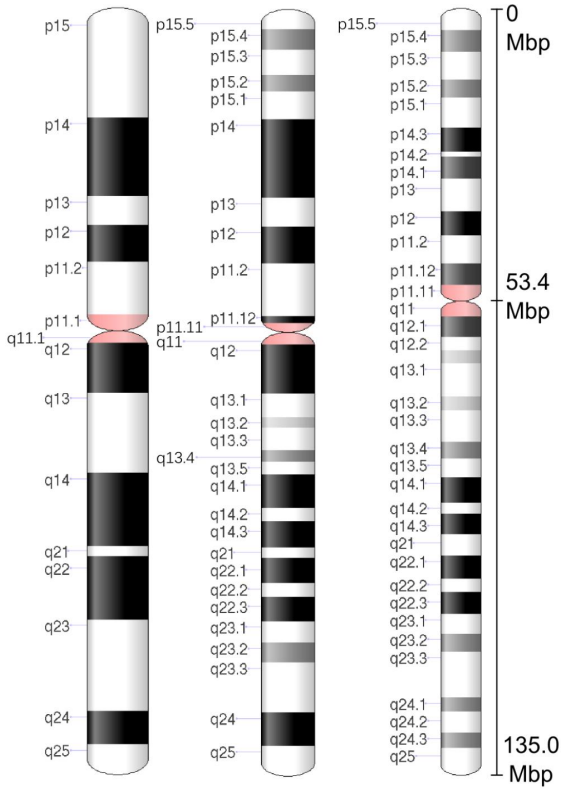

Fig. 2. Ideogram of human chromosome 11.

By comparing chromosomes to ideograms, a laboratory assistant is able to evaluate the overall state of patient's chromosomes and detect abnormalities.

It should be mentioned that there is more than one ideogram for a single chromosome. As it can be seen in Fig. 2, multiple ideograms are introduced to chromosome 11. Multiple ideograms are necessary because chromosome images vary due to chromosome size and quality of laboratory equipment. Also, ideogram images are provided by various vendors.

At the current state of prenatal diagnostic industry, the aforementioned process is conducted manually or in

a partially automated manner. This poses a certain problem, since the karyotyping process is time-consuming and prone to error due to a human factor.

Therefore, there was a strong need for automation of a karyotyping process, and a general algorithm for its automation has been proposed [3, 4]. The proposed algorithm is designed to recognize pathologies by analysing human chromosomes and ideograms. For this purpose, both chromosomes and ideograms have to be converted to a single format. [3] covers feature detecting from chromosome images. However, ideograms have to be converted too.

While having simpler and more predictable structure, ideograms still form a significant dataset, so the process of their recognition is complex enough to consider the possibility of its automation. Also, as it was mentioned before, there can be more than one ideogram for each chromosome.

Another reason to add ideogram recognition is a possibility to use ideograms for storing additional data. For example, a lab assistant would be able to decode and store common chromosome abnormality in a separate ideogram. Since this approach to karyotyping is not a traditional one, this might provide additional benefits to its automation.

This paper considers the problem of the automatic detection of the ideogram feature. The ideogram image would serve as input data, and the result of feature extraction will be represented as data objects.

2. Related works

The problem of the automation of recognising the chromosome pathologies is known and covered in a number of research papers.

[3] offers a general overview of the domain. To summarize, there are proposals of automation algorithms, but there is no comprehensive solution. Moreover, existing algorithms rely on machine learning and datasets which may pose a challenge in the medical domain. Gathering a dataset of medical data can be hard because of privacy issues and the diversity of equipment and technuqies which make medical data heterogenous. Because of this, the general algorithm proposal cannot rely on dataset-driven chromosome recognition.

As an alternative, a comparison of chromosomes and ideograms is proposed, which means feature extraction from chromosomes and ideograms. While the extraction of a chromosome feature is covered by existing research [3, 7-8], extracting data from ideograms has not been widely covered. However, due to relative similarity between chromosomes and ideograms, some known techinques for chromosome feature detection may be applied to ideograms as well.

3. Problem Solution Method

Ideograms were initially introduced as the models of chromosomes, used for visual reference and comparison. By matching ideograms and chromosomes, a lab assistant can determine abnormalities in the chromosome structure.

The main feature utilized in the aforementioned matching process is chromosomal bands [9]. Bands are stripes of specific colour, arranged in a specific pattern. This pattern is used to identify a chromosome (*fig. 3*).

Fig. 3. Side-by-side comparison of ideogram 7 and chromosome 7.

While bands can have a variety of colours, the general proposal [3] requires band classification according to one of two groups: white band of black band. Apart from colour, each band has its length. Therefore, after feature extraction, both chromosome and ideogram will be stored in a similar data structure: an ordered list, where each element represents a band. Each band can be either black or white, and have its length.

Chromosomes and ideograms share common traits, so techniques of extracting the chromosome feature could be reused to a certain extent. For both chromosomes and ideograms, the general idea of the feature extraction is determining the main “axis” of an object and getting colour readings along this axis. After that, normalizing data and thresholding colour into black or white will result in a recognized object. However, normalizing data and removing noise will be different for ideograms and chromosomes.

Figure 4 depicts the general algorithm of feature extraction. The image of ideogram is treated as a matrix where each element (grayscale pixel) is a number from 0 (black) to 255 (white).

The first step of the algorithm is noise removal. It consists of Non-Local Meams denoising [10] and Median filter [11]. After that, an image colour is binarized. According to a test dataset, a threshold $COLOUR\_THRESHOLD = 250$ is established.

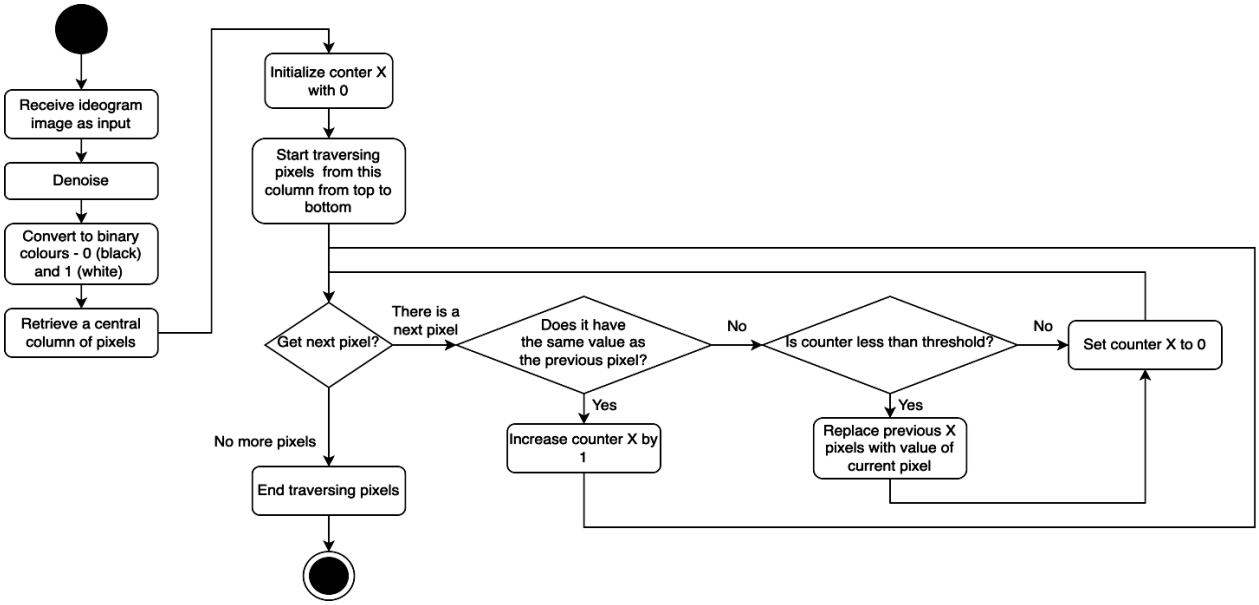

Fig. 4. General algorithm flowchart.

Therefore, having a possible colour range from 0 (black) to 255 (white), any pixel with value less than 250 is converted to black. Any other pixel is converted to white.

As the next step, a central axis of ideogram is retrieved. Due to the established image format, it is considered to be a central vertical column of pixels. At this point, this column is treated as a vector y of n elements containing zeros and ones (equation 1).

$$\vec{y} \in \{0,1\}^n \tag{1}$$

After retrieving the vector, it is necessary to clear it from accidental values. Noise removal partially addressed this issue, but at the end of algorithm execution data should be wellf-formed and non-fuzzy. So, it is prefereable to perform extra data clearing.

To address this task, the following approach is adopted: a pixel threshold *PIXEL_THRESHOLD* = *5* is declared. According to the test dataset, a value of 5 is established for the threshold. A *COUNTER* is initialized with 0.

After this, a vector is traversed. With each element x of same value, *COUNTER* is increased by 1. When the first element x of other value is encountered, the counter is compared to the threshold. If *COUNTER* is less than threshold, vector elements from x – *COUNTER* to x are replaced by a value of x. Thus, short sequences of data are classificated as noise and cleared. Clearing the data concludes the process of feature detection. Figure 5 demonstrates a plot built with the extracted data.

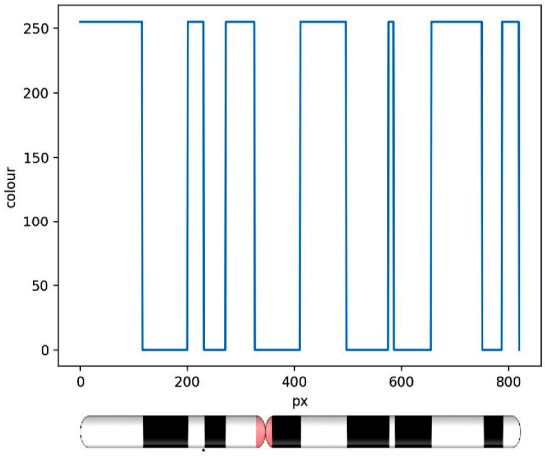

Fig. 5. Extracted ideogram data.

4. Program implementation and verification

A software prototype has been implemented to verify the efficiency of the aforementioned algorithm. The prototype is built with the use of Python programming language and OpenCV library. Figure 6 is a plot built using a PyPlot library.

For verification, a test set containing 10 ideograms from different sources and representing different chromosomes has been considered. Six of them were processed successfully, and processing of remaining 4 ideograms yielded partial success. Major reasons for faulty feature detections are low resolution of images and noise that cannot be automatically removed. However, ideograms of high resolution, like those depicted in Fig. 2, are processed successfully.

5. Conclusions

In the paper, the task of detection and extraction of ideogram features has been considered. Ideogram feature extraction is a partial solution of a bigger problem – an algorithm for detection of chromosome pathology.

The review and analysis of related works have been conducted. While processing images to obtain chromosomal data is not a novel problem, extracting data from ideograms is not covered in the existing papers. However, some techniques for recognition of the chromosomal features were reused – for example, using central axis to obtain image colour profile.

The algorithm has been proposed and verified by implementing the software prototype. As a result of using diverse test dataset, the success rate of the proposed method is 60%. Further research could be directed to cover more formats of ideogram images. Low resolution ideograms might be recognized successfully after image upscale, and noisy images might be handled with the use of adaptive thresholding.

References

[1] C. O'Connor, "Chromosome mapping: Idiograms", *https://www.nature.com/scitable/topicpage/chromosome-mapping-idiograms-302/*, 2008.

[2] C. O'Connor, "Karyotyping for Chromosomal Abnormalities", *https://www.nature.com/scitable/topicpage/karyotyping-for-chromosomal-abnormalities-298/*, 2008.

[3] O. Pysarchuk and Y. Mironov, "Chromosome Feature Extraction and Ideogram-Powered Chromosome Categorization", *Advances in Computer Science for Engineering and Education. Lecture Notes on Data Engineering and Communications Technologies*, vol 134, pp 427–436, Springer, Cham. 2022.

[4] O. Pysarchuk, Y. Mironov, "Decision support system for medical pathology recognition", *Science-Based Technologies*, vol. 49 No. 1, pp. 13-22, 2021. (Ukrainian)

[5] S. Moorthie, et al, "Congenital Disorders Expert Group. Chromosomal disorders: estimating baseline birth prevalence and pregnancy outcomes worldwide", *Journal of Community Genetics*, vol. 9(4), pp. 377-386, 2018.

[6] X. Zhang, et al, "Cytogenetic Analysis of the Products of Conception After Spontaneous Abortion in the First Trimester", *Cytogenetic and Genome Research*, vol. 161, pp. 120-131, 2021.

[7] R. Nandakumar and KB Jayanthi, "Feature Extraction for the Classification of Human Chromosomes from G-Band Images using Wavelets", *International Journal of Engineering Research & Technology (IJERT) ICEECT,* no 8(17), pp. 67-72, 2020.

[8] M. Moradi and K. Setarehdan, "New features for automatic classification of human chromosomes: A feasibility study", *Pattern Recognition Letters*, vol. 27(1), pp. 19-28, 2006.

[9] S. Kumar, A. Kiso, and N. Abenthung Kithan, "Chromosome Banding and Mechanism of Chromosome Aberrations," Cytogenetics - Classical and Molecular Strategies for Analysing Heredity Material, Jul. 2021.

[10] A. Buades, B. Coll, J. Morel, "Non-Local Means Denoising", *IPOL Journal*, vol. 1, pp. 208-212, 2021.

[11] "Median Filter", https://en.wikipedia.org/wiki/Median_filter, Nov 21, 2022.

ВИДІЛЕННЯ ОЗНАК ІДЕОГРАМ ДЛЯ РОЗПІЗНАВАННЯ ХРОМОСОМНИХ АНОМАЛІЙ

Олексій Писарчук, Юрій Міронов

В даній публікації запропоновано підхід до розпізнавання зображень з хромосомними ідеограмами. Ідеограма – відображення здорової хромосоми, яке використовується в процесі каріотипування – процедурі, розробленій для діагностування хромосомних аномалій.

Розпізнавання ідеограм – це частина загального алгоритму з діагностування хромосомних аномалій. Згідно з даним алгоритмом, хромосоми та ідеограми мають бути перетворені в єдиний формат даних для подальшого порівняння.

Алгоритм розпізнавання ідеограм, що його запропоновано в даній публікації, приймає на вхід зображення ідеограми та повертає структуру даних, яка містить властивості ідеограми. Для підтвердження результативності алгоритму було розроблено програмний прототип.

Yurii Mironov. PhD student at Software Engineering Department of National Aviation University - Kiev, Ukraine.

Scientific interests include computer vision, object recognition, modeling, software architecture, model-driven development.

Oleksii Pysarchuk. Doctor of Science in Engineering Sciences; Professor of Computer Engineering Department at Igor Sikorsky Kyiv Polytechnic Institute, Ukraine.

Scientific interests include systems analysis, modeling, optimization and information security.