Identification of column edges of DNA fragments by using K-means clustering and mean algorithm on lane histograms of DNA agarose gel electrophoresis images

Muhammed Kamil Turan[a], Eftâl Şehirli[a], Abdullah Elen[a], İsmail Rakıp Karaş[a]
[a]Karabuk University (Turkey)

ABSTRACT

Gel electrophoresis (GE) is one of the most used method to separate DNA, RNA, protein molecules according to size, weight and quantity parameters in many areas such as genetics, molecular biology, biochemistry, microbiology. The main way to separate each molecule is to find borders of each molecule fragment. This paper presents a software application that show columns edges of DNA fragments in 3 steps. In the first step the application obtains lane histograms of agarose gel electrophoresis images by doing projection based on x-axis. In the second step, it utilizes k-means clustering algorithm to classify point values of lane histogram such as left side values, right side values and undesired values. In the third step, column edges of DNA fragments is shown by using mean algorithm and mathematical processes to separate DNA fragments from the background in a fully automated way. In addition to this, the application presents locations of DNA fragments and how many DNA fragments exist on images captured by a scientific camera.

Keywords: Lane histogram, k-means clustering algorithm, mean algorithm, DNA fragments

1. INTRODUCTION

Gel electrophoresis is a method to separate molecules such as DNA, RNA, and protein and to identify fragments of them. Agarose gels have already been contented with forming suitable sizes for DNA fragments and being used to make them walk[1,2]. Ethidium bromide is preferred to stain DNA to be able to show the fragments of DNA molecules in agarose gel[3,4]. After applying electrophoresis process on agarose gel, DNA walks from anode to cathode in an agarose gel. While DNA molecules are walking on agarose gel, it draws fragments and owing to ethidium bromide, each fragment can be visualized on agarose gel[4]. After that, images of agarose gels which includes DNA molecules as a fragment shapes can be captured by using scientific cameras.

It is very important to detect DNA fragments existing on captured images by a software application. Since calculations of size, weight, quantity of DNA fragments and analysis on each DNA fragment are possible to realize thanks to applications that are able to do these tasks in a short time.

There have been several similar studies about detecting column edges of DNA fragments. S.C. Park and his friends used k-means clustering algorithm and conducted subsequent local image processing at their studies. After obtaining lane histogram, they estimated average of lane width values with k-means clustering algorithm. Then they made the whole image partition into small portions for subsequent local image processing to track the center and detect the partitioned lanes[5]. Skutkova and friends made an image enhancements. They found the first pixels of each lane by calculating intensity mean values. They made lane tracking starting with first pixel of each lane column. Therefore, they obtained each DNA fragments on agarose gel electrophoresis images[6].

The purpose of this study is to present an algorithm for software applications to identify DNA fragments on captured images in a fully automated way. It was realized by using k-means clustering and mean algorithm together on obtained lane histogram of different types of agarose gel electrophoresis images. In addition to this, it is also aimed at preparing a ground to do projection based on y-axis.

2. MATERIALS AND METHODS

2.1. Materials

In this study, agarose gel electrophoresis images which were necessary to be able to develop an application were captured during experiment in micro geneticand molecular biology laboratory of Karabuk University Training and Research Hospital. Images are at the different formats and sizes. The scientific camera whose technical properties is shown below in Table 1 was used to capture images.

Parameters	**Values**
Image Sensor	½ inch, colorful 1.4MP CCD
Frame Rate	12.5fps@1360x1024, 15fps@680x520
Exposure Time	0.1ms~60mins
Color Filter	R, G, B
Scan Mode	Progressive Scan
Dynamic Range	60dB

Table 1. Technical properties of scientific camera.

Some of the captured images which were used in this study are shown in Figure 1 below.

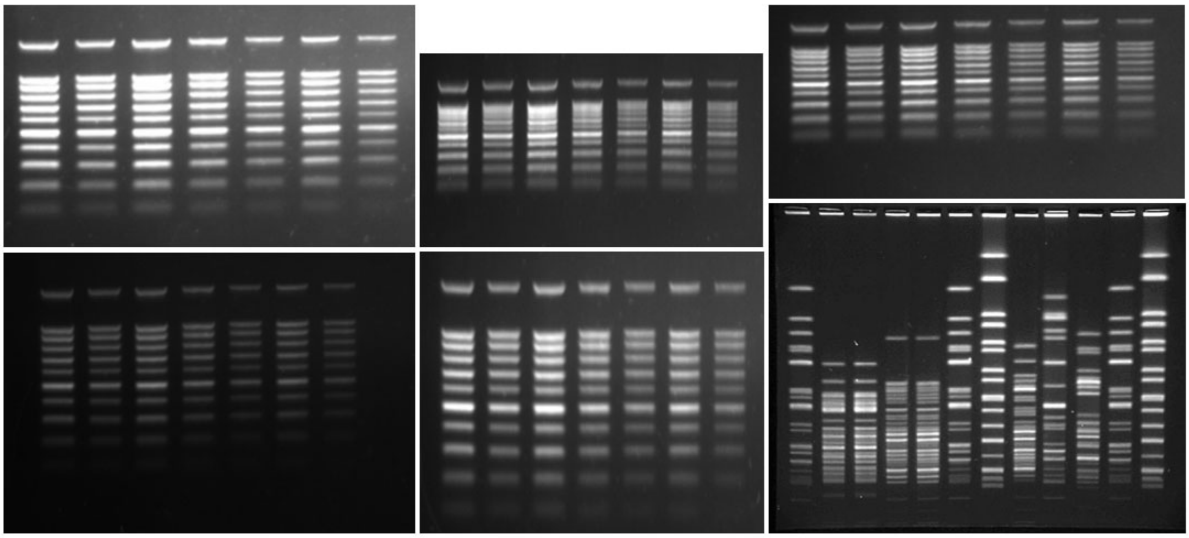

Figure 1. Some of the captured images which were used in this study.

Captured images were tested in the developed software application in this study. The software application is a windows desktop application. Visual Studio development environment was selected to develop the software application. It was written in C# programming language. Some image processing dynamic link libraries of Emgu CV that are open source files were preferred to show lane histograms of agarose gel electrophoresis images.

2.2. Methods

In this study, k-means algorithm and mean algorithm were applied on lane histograms of captured images and detection process of hills existing on lane histograms of images was realized. The schematic representation of processes applied on this study is shown in Figure 2 below.

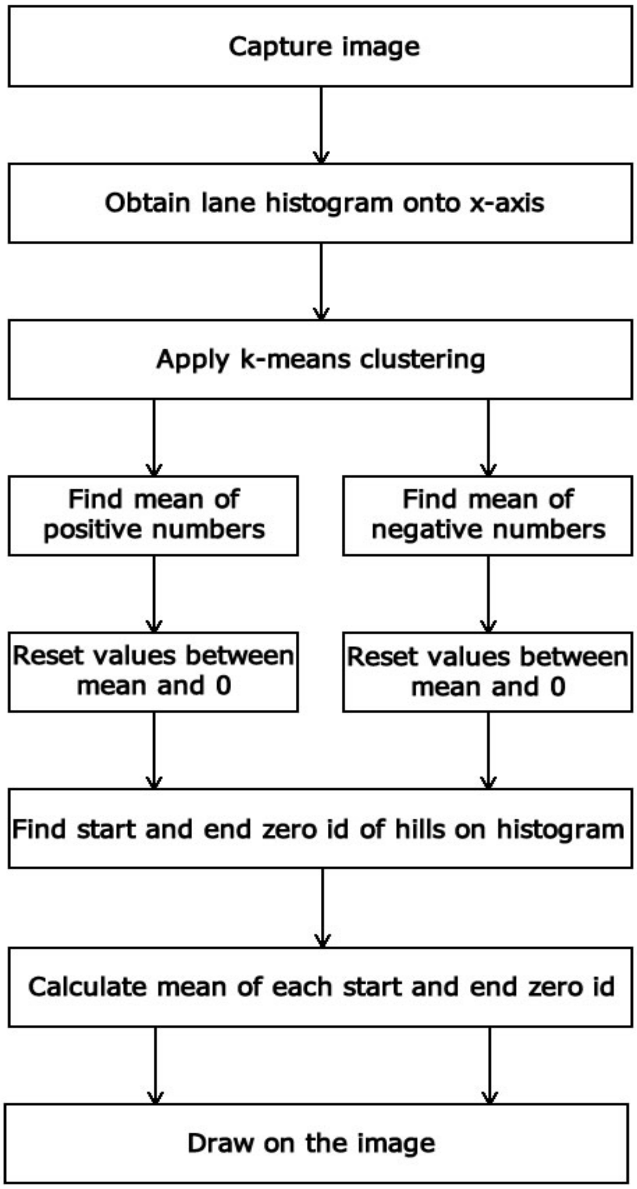
Capture image
Obtain lane histogram onto x-axis
Apply k-means clustering
Find mean of positive numbers
Find mean of negative numbers
Reset values between mean and 0
Reset values between mean and 0
Find start and end zero id of hills on histogram
Calculate mean of each start and end zero id
Draw on the image

Figure 2. Schematic representation of the algorithm.

First of all, lane histogram onto x-axis of loaded image was obtained. Lane histogram onto x-axis is a type of histogram that obtains pixel intensities according to x-coordinate information. In other words, while x coordinate value is constant, y coordinate value changes from 0 to image height. Each intensity value of (x,y) point whose x values are constant and y values change is added. Therefore, images are transformed from 2D to 1D. A lane histogram of an agarose gel electrophoresis image is shown in Figure 3 below.

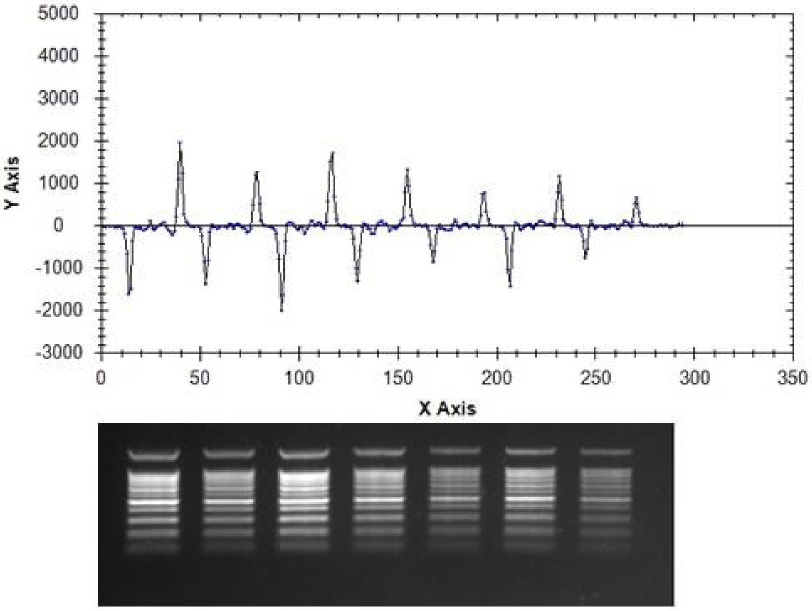

Figure 3. Lane histogram onto x-axis of a sample image.

After obtaining lane histogram onto x-axis, k-means clustering algorithm was applied to classify point values of lane histogram into three parts such as left side values, right side values and undesired values shown in Figure 4. Hence, k value is selected as three. In order to identify DNA fragments on images, left side values and right side values are necessary; however values defined as undesired values are not necessary in order to show DNA fragments.

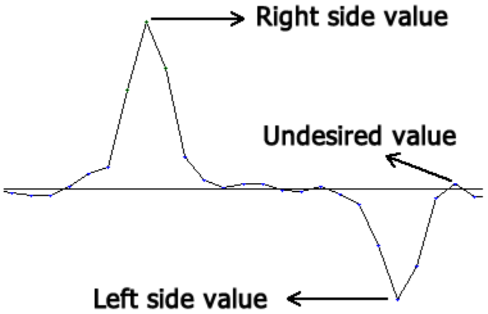

Figure 4. Representation of left side, right side and undesired value on lane histogram.

Undesired values are not necessary to identify DNA fragments because left side and right side values represent column edges of DNA fragments. After classifying values in 3 different classes, mean algorithm was used to eliminate undesired values on the lane histogram. But, mean algorithm must separately be applied for both positive values and negative values of the lane histogram since mean value becomes almost zero if positive values and negative values are not separated. As a result of that, undesired values are not eliminated. After mean algorithm was applied for both sides of the lane histogram x-axis, values between mean value of positive values and 0 were changed with 0. Similarly, values between mean value of negative values and 0 were changed with 0. As a result, obtained result for the sample captured image is shown in Figure 5 below.

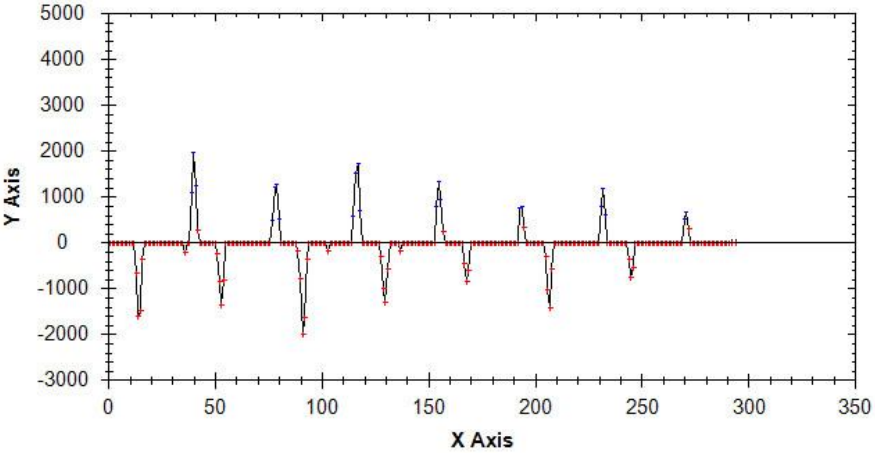

Figure 5. Lane histogram after applying mean algorithm for sample captured image.

After that, lane histogram was examined along x-axis to identify places of zero values. Because peak values of hills on the lane histogram are wanted to be shown on captured images as column edges of DNA fragments, it is important to find where zero values start and end. After finding places of zero values, mean values are calculated between start points and end points of zero values. Negative peak values of hills on lane histogram are left column edges of DNA fragments and positive peak values of hills on lane histogram are right column edges of DNA fragments. Moreover, the number of the positive peak values of hills or the negative peak values of hills are the number of DNA fragments of captured images. As a result, column edges of DNA fragments can be drawn on the captured images and the number of the DNA fragments can be written in the software application.

3. RESULTS

Images that have different sizes, different backgrounds, include different numbers and sizes of DNA fragments were captured during experiment. In the software application, these captured images were used and tested. Obtained results of the images which are shared in the Section 2 are shown in Figure 6 below. If captured images is not cropped, the column edges of images are also drawn at the left and right sides of images. To prevent this, images are cropped or these column edges that are the most left edge and the most right edge is not drawn automatically by assuming that they are not any column edges of DNA fragments.

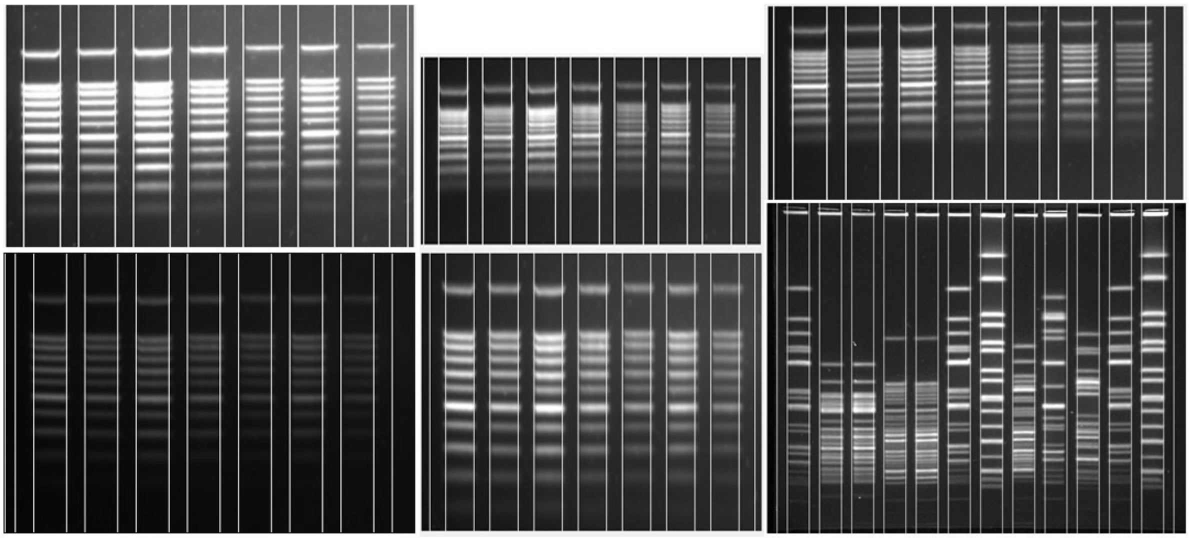

Figure 6. Result images of the each type of captured images.

4. DISCUSSION AND CONCLUSION

In this paper, a fully automated application is proposed that a high-accurately identifies column edges of DNA fragments on many types of agarose gel electrophoresis images. The purpose is both to detect DNA fragments based on x-axis and to be able to develop a method to do projection based on y axis to obtain only one single rectangular part inside a DNA fragment as a future work. In this study, detection column edges of DNA fragments based on x-axis was realized by using many types of agarose gel electrophoresis images successfully. Furthermore, all DNA fragments in agarose gel electrophoresis images was separated with each other. As a future work, it is thought to obtain one single rectangular part of each DNA fragments in the application. Therefore, users will make observations on agarose gel electrophoresis images in a detailed way by using both DNA fragments and one single rectangular part.

5. ACKNOWLEDGEMENT

We desire to thank ICDIP to give us a chance to introduce our algorithm and developed software application and present it on the conference.

We would like to appreciate Karabuk University Training and Research Hospital to be able to utilize their equipment and laboratory.

This study was supported by TUBITAK - The Scientific and Technological Research Council of Turkey (Project No: 112Y050) research grant. We are indebted for its financial support.

6. REFERENCES

[1] Borst, P., “Ethidium DNA Agarose Gel Electrophoresis: How it started”, IUBMB Life, 57(11), 745-747 (2005).

[2] Caridade, C. R., Margal, A. S., Mendonga, T., Pessoa A. M., Pereira S., “An automatic Method to identify and extract information of DNA bands in Gel Electrophoresis images”, 31st Annual International Conference of the IEEE EMBS, 1024-1027 (2009).

[3] Diaz F., Bayona-Bafaluy M. P., Rana M., Mora M., Hao H., Moraes C.T., “Human mitochondrial DNA with large deletions repopulates organelles faster than full-length genomes under relaxed copy number control”, Nucleic Acids Res., 30(21), 4626–4633 (2002).

[4] Kasap M., Torol S., Gacar G., Budak F., “Ethidium Bromide Spot Test Is a Simple Yet Highly Accurate Method in Determining DNA Concentration”, Turk J Med Sci 36(6), 383-386 (2013).

[5] Park S. C., Na I, S., Han T. H., Kim S. H., Lee G. S., “Lane detection and tracking in PCR gel electrophoresis images”, Computers and Electronics in Agriculture, 83, 85-91 (2012).

[6] Skutkova H., Vitek M., Krizkova S., Kizek R., Provaznik I., “Preprocessing and Classification of Electrophoresis Gel Images Using Dynamic Time Warping”, International Journal of Electrochemical Science 8, 1609-1622 (2013).