Implementation of Digital Image Processing to Electrophoresis Image for Identification of DNA Bands with OpenCV in Python

Naila Fadhilah Fithriah[1], Elvayandri Muchtar[2], Waskita Adijarto[3]

School of Electrical Engineering and Informatics, Bandung Institute of Technology
Jalan Ganeca 10, Bandung, West Java, Indonesia
[1]nailafadhilah1998@gmail.com, [2]eva@elka.ee.itb.ac.id, [3] waskita@ee.itb.ac.id

***Abstract*— Electrophoresis is a process of electrochemical breakdown of DNA bands in which the RNA / DNA protein fragments are made to move in a gel medium that is flowed by a current. Gel Electrophoresis itself is one of the most commonly used techniques due to the wide field of use of this technique. Digital image processing and pattern recognition techniques can be used to extract qualitative and quantitative information from an image. In addition, digital image processing can overcome human errors that may occur from manual analysis. This article will explain the implementation of digital image processing from the results of electrophoretic images to identify protein sequences in DNA bands using OpenCV in Python. The process of digital image processing from electrophoresis is carried out by thresholding, Contrast Limited AHE (CLAHE), morphology, searching for the contour and midpoint of the band, and comparing the bands to the reference bands in the lane ladder.**

***Keywords*— Electrophoresis, OpenCV, Python.**

I. INTRODUCTION

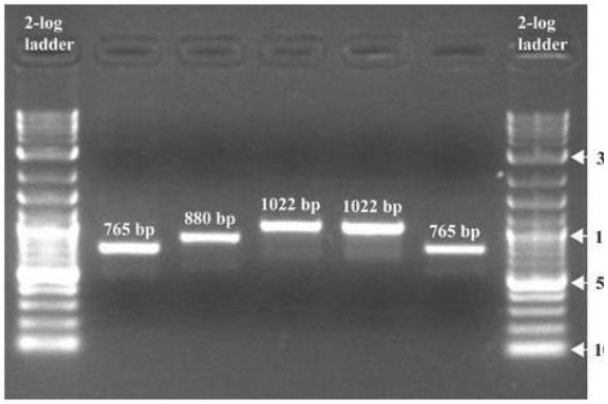


Figure 1 Electrophoresis Image Visualization

The development of gel electrophoresis as a method of DNA separation and analysis has made significant progress and influenced advances in the field of molecular biology over the past 20 years. Electrophoresis is a method for separating molecules based on their size, electric charge, and other physical properties. DNA gel electrophoresis refers to a technique in which DNA macromolecules are forced across a variety of gels, which are colloids in solid form by means of an electric current. Two types of matrix gel, agarose gel and polyacrylamide, are commonly used for DNA gel electrophoresis. In this gel, the electrophoretic mobility of macromolecules is determined by the volume fraction of the pores in the gel that are accessible to the macromolecules. Therefore, a mixture of DNA molecules of different sizes will separate into discrete clumps (called bands) in the electrophoresis process. Scientists use electrophoresis to obtain information about the substance under study, such as comparing the composition of a sample, or calculating the number and properties of various elements present in a sample set.

In general, images with biological elements are a major challenge for computer interpretation because of their irregular and varied shapes. The images to be analyzed in biometry can come from a microscope, medical scanning system, electrophoresis, remote sensing, or simply from photographing an illuminated object. With advances in image analysis techniques in recent years, computers have become an important tool for scientists and professionals in data acquisition and interpretation. Digital gel electrophoresis analysis is emerging as one of the most important applications, for reducing human error and increasing the speed of data evaluation. In order to accurately and quantitatively analyze gel electrophoresis, the resulting image contains bands or dots that reflect the number and characteristics of each component. Digitizing gel images provides a way to objectively measure information. Image processing and analysis techniques convert the information carried by gel electrophoresis into reliable, accurate, neat, and consistent numerical, graph, table, and visual formats.

In this article, the authors implement digital image processing for electrophoretic images using OpenCV in Python. The use of OpenCV is determined because of the ability of OpenCV which already has built-in image processing functions such as binary thresholding, histogram transformation (both linear and non-linear filtering) to increase the parameters of the raw image before processing. Mathematical morphological methods such as opening, closing, Top-Hat, and Hitor-Miss transformations are used to measure the morphological features of objects; measurements such as length, area, histogram, etc. extracted and interpreted using stereology, form statistics, or classification methods.

II. System Design

Implementation of digital image processing results from electrophoresis for identification of DNA bands with OpenCV in Python is done only with software. The design of a digital image processing system is done by first making a flow chart that explains what each part of the implementation program is doing to get the results of identifying digital images from an electrophoretic image.

Here is a flow chart of the system.

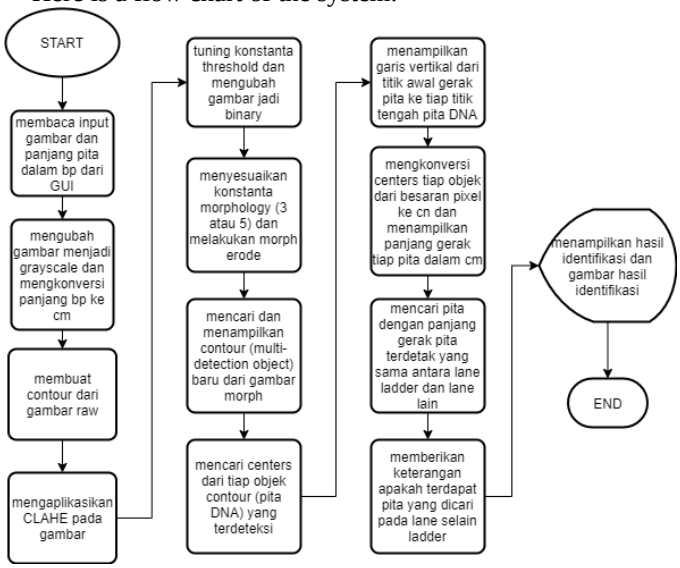


Figure 2 Electrophoresis Image Identification Flowchart

The concept used in processing electrophoretic digital image results is to use thresholding and CLAHE, where the process is carried out aimed at reducing noise in the image acquired by electrophoresis so that multiple object detection can be carried out which can search for the contours of each DNA sequence band so that the center point of each DNA sequence is obtained. tape. The center point of each band is then measured from the starting point of the motion of the tape, and compared with the ladder tape. If the base length of the DNA sequence is the same as the ladder band, identification is successful.

The first process that is carried out is to read the image from electrophoresis and convert it to grayscale. Then, CLAHE was done. CLAHE is an improved writing algorithm from AHE. CLAHE is a generalization of AHE (Adaptive Histogram Equalization). In contrast to HE (Histogram Equalization) which operates in the entire region in the image, CLAHE operates in a small region in a grayscale image called a tile. CLAHE application is done to increase the contrast in the image by providing a clip limit which states the maximum height limit of a histogram.

$$\beta = \frac{M}{N}\left(1 + \frac{\alpha}{100}(S_{max} - 1)\right)$$

The above equation is the CLAHE equation. The variable M is the area of the image portion, N represents the grayscale value (256), and α is the clip factor which represents the increase in the limit of a histogram which is between 0 and 100.

A histogram that is above the clip limit value is considered to be an excess of pixels (excess) which will be distributed to the area around the bottom of the clip limit so that the histogram is even. The following is an illustration of the excess pixel distribution.

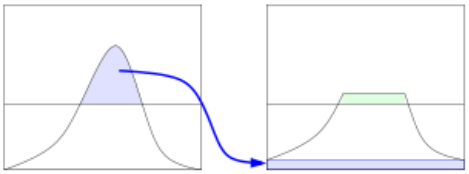


Figure 3 Histogram Excess Pixel Distribution

After obtaining the CLAHE result image, the CLAHE result image is reprocessed by the thresholding method. The thresholding method is the simplest image segmentation method. The application is to separate the part of the image that you want to analyze and distinguish between the pixels you want to process and those that don't.

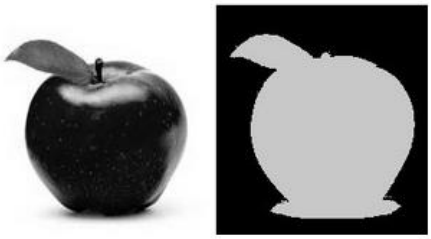


Figure 4 Threshold Image Visualization

Implementation of the program is carried out by utilizing a binary threshold, where the functions are as follows.

$$dst(x,y) = \begin{cases} maxVal & if\ src(x,y) > thresh \\ 0 & otherwise \end{cases}$$

When the src (x, y) pixel intensity is greater than the threshold point, the pixel intensity is changed to MaxVal. Also, the pixel intensity is set to zero.

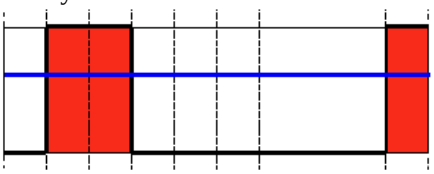


Figure 5 Threshold Binary Method

After that, the image will be processed by morphological transformation. The morphological transformation used is erosion (ERODE). The concept of erosion morphological transformation is the erosion of the boundaries of the background object and keeping the main object white. The pixel in the image only remains 1 if all pixels under the kernel are 1, otherwise it will be degraded (made zero). The kernel size will determine how many pixels near the money limit will be discarded. The thickness of the white object will decrease as the kernel size increases. This section will remove white noise in the image and remove two connected objects due to deficiencies in the image.

Figure 6 Erode Morphology Visualization

From the image that has been optimized through the filtering processes above, a contour search of the DNA band in the image can be done. A contour is a curve that connects all finite continuous points of the same color or intensity. Contours are very useful for shape analysis and object detection and recognition.

After the contours of each DNA band are found, a midpoint search can be performed. The midpoint search is performed using the moments function in OpenCV. Moments will count all the moments up to the three polygon and raster shapes. Then, we get the midpoints of the DNA band. Centers or DNA band points on a lane other than the lane ladder are then sequenced and then compared with the DNA ladder. The following is a snippet of a program for sequencing the midpoint of the detected DNA band.

```
centres_np = np.array(centres)
        centres_sorted =
centres_np[centres_np[:,0].argsort()[::-1]]
        centres_sorted

        dna_lanes = centres_sorted[0:6]
        reference_lanes = centres_sorted[6:]

        for center in dna_lanes:
            closest_reference_idx =
(np.abs(reference_lanes[:,1]-
center[1])).argmin()

        dna_lanes = centres_sorted[0:6]
        reference_lanes = centres_sorted[6:]

        user_input_example = [0, 120]
        closest_reference_idx =
(np.abs(reference_lanes[:,1]-
user_input_example[1])).argmin()
        reference_center =
reference_lanes[closest_reference_idx]

        max_delta_px = 20
        ket = []

        for center in dna_lanes:
            if np.abs(reference_center[1] -
center[1]) <= max_delta_px:
                ket.append("Yes")
            else:
                ket.append("No")
```

The words “Yes” and “No” at the end indicate whether the DNA trait you are looking for on the DNA band on the lane other than the lane ladder (reference) has been found or not..

III. Implementation Result

In the implementation of digital image processing results from electrophoresis for identification of DNA bands with OpenCV in Python, application and testing are carried out using the Raspberry Pi 4 control unit. Graphical User Interface is made on the Local Web Server where the back-end side of this web application is made in Python, so that it is in line with the use of OpenCV in Python for the implementation of this part of digital image processing.

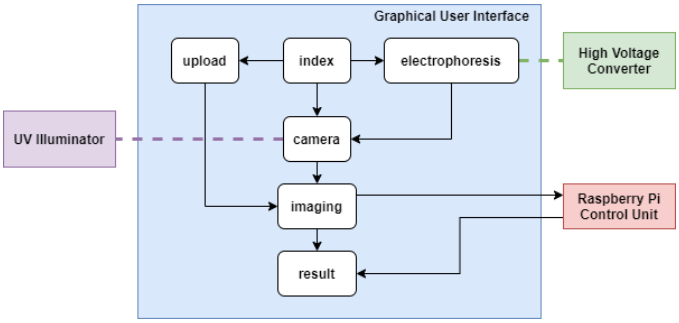


Figure 7 Identification Device UI Diagram (Imaging is processed in Raspberry Pi Control Unit)

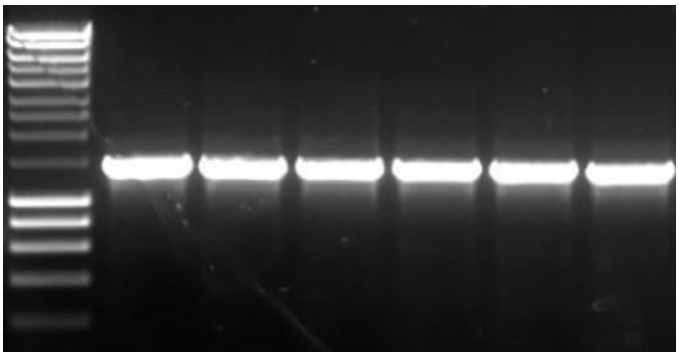
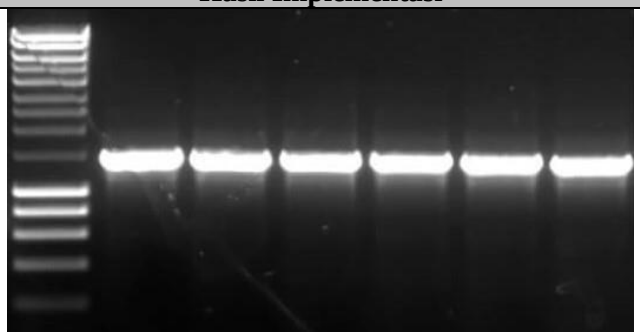
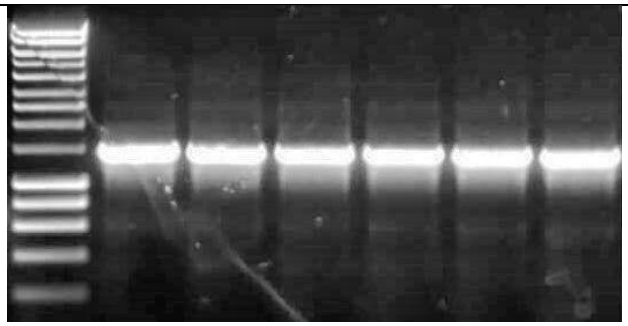


Figure 8 Original Image before Image Processing

Following are the results of implementing a digital image processing program for the following electrophoretic image results.

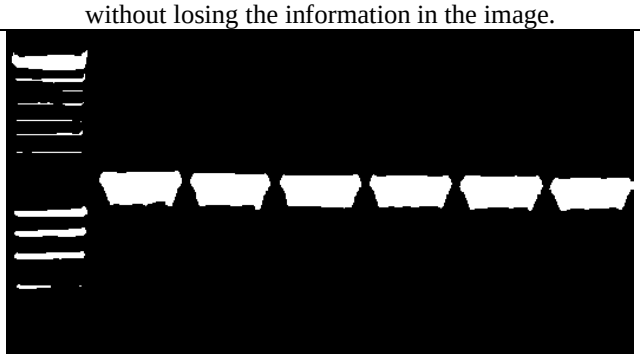
Table 1 Gel Electrophoresis Image Processing Implementation Result

Hasil Implementasi

The original image is converted to grayscale at the bit rate. Changing the image value to grayscale is done to produce a better image when passing through other filtering methods.



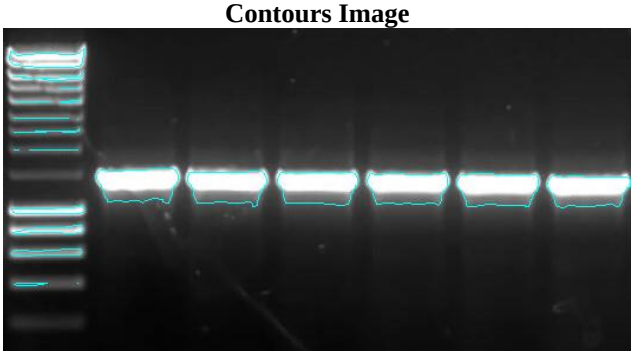
Results of images that have gone through CLAHE (histogram equalization). The contrast of the previously low image increases and shows parts of the image that were previously covered with noise due to image resolution.

The results of the CLAHE image are then performed binary thresholding to convert the image into a binary image without losing the information in the image.



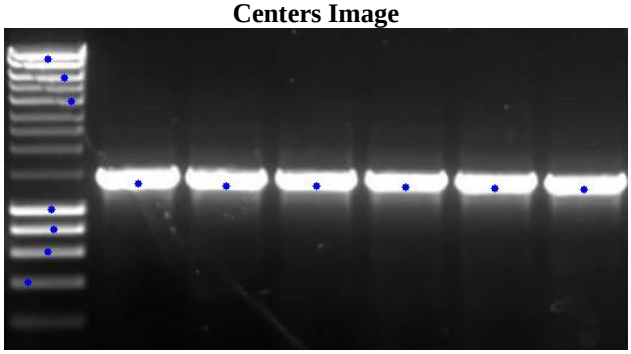
The results of the threshold image are then subjected one again to erode morphology to ensure that the main bands (identification bands other than ladder bands) are not connected and form the same object.

Contours Image

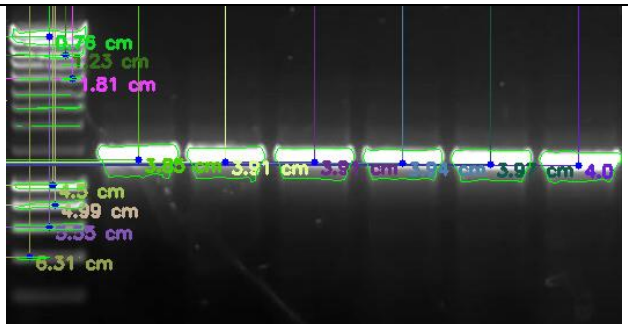


Showing the results of the contour of the image that has been morphologically eroded before, showing that the bands in the ladder 2-6 are well separated.

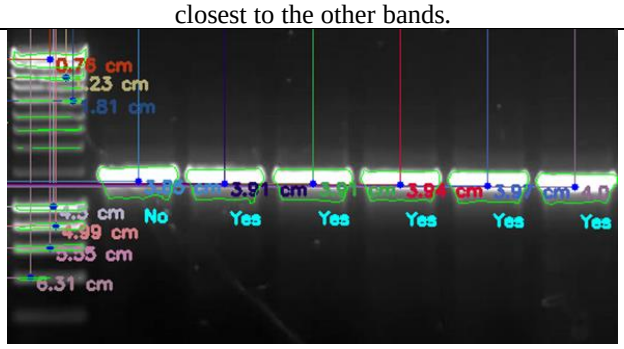
Centers Image



Then, the results of the image that have been morphologized are searched for the contours and the midpoint of each contour is obtained and the results are as above.



The resulting midpoint and contour drawing are put together. Then, a vertical and horizontal line is drawn from the starting point of the DNA band motion. Assuming the comparison of the length of the DNA band base with the distance from the starting point to the end point of the band movement, a linear 8 cm distance conversion is equal to the maximum band length of 0-2000 bp. The height of motion of the tape is obtained in centimeters. This value is converted to pixels which will then be compared with the bands in the lane ladder (lane 1) and find which band is closest to the other bands.



If the delta of the difference between the distance of the band that determines the quality of rice in the lane ladder is the same as the distance of the band motion on the lane other than that with a very small difference, it can be concluded that the band that determines the DNA properties sought has been found in the lane section containing the identified DNA band.

From the results obtained, it can be seen that the implementation of digital image processing results from electrophoresis for identification of DNA bands with OpenCV in Python has been successful. The level of accuracy of the

implementation of this system cannot be tested because it has not carried out direct testing in the laboratory. Suggestions for improvement in the future are to create other implementation programs that can be compared with existing methods and / or improve the success of digital image processing from electrophoretic images.

IV. Conclusion

Digital image processing results of electrophoresis for identification of DNA bands is implemented using OpenCV in Python. The implementation process is done by changing the image first to black and white, then thresholding the image. The resulting threshold image is reprocessed with CLAHE to reduce noise to the maximum. Then, the image will be converted back into a binary image. Different tuning and parameters are needed for each image to be processed because the digital image processing, especially with thresholding, is strongly influenced by light intensity, shooting distance, and image sharpness. After obtaining the binary image, the image is processed with the morphology function in Python to improve image quality prior to a multiple-object detection search using the find contours method. After the contours of each band are found, the center point of each band can be found. The center points of each band will be compared with the center points of the tape on the lane ladder. If the presence of DNA bands on other pathways besides the lane ladder are parallel to the sequence of reference DNA bands from the lane ladder, it can be concluded that the quality of the DNA sought has been identified. The optimization of the program for the identification of rice DNA band sequences can be done by applying 2D-filtering algorithms such as 2D edge detectors.

V. References

[1] *Agarose* Gel Electrophoresis. https://www.addgene.org/protocols/gel-electrophoresis/ [Accessed on 10 December 2019].

[2] Xian gyun Y., Ching Y. S., Cheriet M., Eugenia W., "A Recent Development in Image Analysis of Electrophoresis Gels", Vision Interface, Canada, pp, 432-438, 1999.: https://www.etsmtl.ca/ETS/media/ImagesETS/Labo/LIVIA/Publications/1999/Ye_VI99.pdf [Accessed on 8 March 2020]

[3] Adiga P.S.U., A. Bhomra, "Automatic analysis of agarose gel images," BioInformatics 17(11), pp. 1084-1090, 2001. https://www.researchgate.net/publication/11630566_Automatic_analysis_of_agarose_gel_images [Accessed on 8 March 2020]

[4] Naima Kaabouch, Richard R. Schultz, Barry Milavetz, An Analysis System for DNA Gel Electrophoresis Images Based on Automatic Thresholding and Enhancement, https://arxiv.org/ftp/arxiv/papers/1607/1607.00589.pdf, [Accessed on 21 April 2020]

[5] OpenCV Documentation, https://docs.opencv.org/, [Accessed on 8 September 2020]