Automated Image Analysis for DNA Fingerprint Gel Electrophrisis Images

Israa Ismail, Ghada S. Eltaweel, Hamed Nassar
Dept. of Computer Sciences, Faculty of Computers & Informatics,
Suez Canal University, Ismailia, Egypt.
{Israa, Ghada_eltawel, nassar}@ci.suez.edu.eg

Abstract

Genetics and genetic analysis experiments can be improved by the application of signal processing techniques in general and image processing techniques in particular. Most of the experiments in genetics produce digital signal or image pattern, which is subjectively analyzed by geneticists. Considering a large number of data to be processed for analyzing any single experiment, it is prudent to develop automatic techniques that could increase throughput and reproducibility of the result. Advances in the image processing field can be effectively applied to achieve this. This paper presents a scheme that aims to detect and segment the lanes in gel electrophoresis DNA images without any human interventions.

Keywords: *Gel electrophoresis (GE), matched filter, watershed segmentation algorithm.*

1. Introduction

Gel electrophoresis is a tool for gene and genomic analysis. The technique has been extensively used for gene identification, isolation and purification. Gel electrophoresis technique produces images consist of several lanes with a lane corresponds to a sample. Each lane has numerous bands and these bands contain valuable information to scientists. The information lies in the position of each horizontal band in a lane. The origin of a sample can be determined by comparing it to a known sample. Two samples are considered to have the same origin if they produce the same pattern of the strand [1]. The process of extracting this information manually consumes a lot of time and repetition. In the worst scenario, the quality of the image produced is bad and inspection cannot be done. Thus in this case no information can be extracted and these lead to a waste of time and cost. In addition, scientists have spent a lot of time to prepare the samples in the process and no results are acquired. The most efficient approach for improving these circumstances is implementing all the tasks automatically. Automated analysis and detection can overcome these drawbacks.

Previous work regarding this problem can be found in [2]–[7], the semi-automatic lane detection method of Elder and Southern (ES) [2] is based on equi-spaced lanes with constant width, where the center of the first and the last lanes in the gel image are manually specified and the number of lanes between the first and the last lanes is given by the user. Kaabouch, N. Et al [3] proposed an algorithm that consists of four main steps: automatic thresholding, shifting, filtering and data processing. They use the automatic thresholding to equalize the gray values of the gel electrophoresis image background.

Akbari A. Et al [4] presented an effective noise filtering technique that a noise model for the description of the noise type is required.

Lin, C.Y et al [5] designed a computerized method to compare the lanes and identify the identical ones. This method segments the lanes and bands in the GE images. In order to describe the position of the bands, they introduce a position vector normalization technique. Then the compared lanes become equivalent to the position vectors. As a result, this method could accurately identify identical lanes.

Cheng, W.Z. Et al [6] presents a method whereby lanes in a GE image are first segmented and converted into a chain code representation. The lane comparison is performed by calculating the longest common subsequence (LCS) in two chain codes.

Akbari, A. Et al [7] present the (ES) semi-automatic lane detection method, iterative moving average filter (IMA) and continuous wavelet transform (CWT) followed by two new methods for lane separation.

Many factors affect the image quality and the patterns in the lanes, such as the applied voltage, field strength, pulse time, reorientation angle, agarose type, concentration, and buffer chamber temperature [8]. In electrophoresis, DNA or other charged molecules are forced to move through the maze formed by the polymers. The mobility is guided by two factors, the mass and the shape of the molecules. The smaller the mass, the faster it moves. As the samples move farther away from the original starting point, the effects of the shape on the molecules start to appear and the bands become blurry.

In this paper, we present a scheme that involves the pre-processing, lane detection and lane segmentation for gel electrophoresis DNA images.

2. Material and Methods

The proposed scheme consists of the pre-processing, lanes detection and lanes segmentation. After the image acquisition the proposed scheme starts by converting the image into a grayscale image, then background removal and enhancement of the gel image, The main purpose of the preprocessing stage is to enhance the image and make it sharper, The lanes detection and segmentation used the intensity profile to detect lanes and matched filter to enhance the bands' shape, then the watershed segmentation algorithm is applied which actually segments the lanes. Figure 1 shows the block diagram for the proposed scheme.

2.1 The Preprocessing stage

The noise present in the image is due to improper gel preparation, image acquisition setup and instrument noise. We have used contrast enhancement and Un-sharp masking for reducing the noise and enhancing the gel image.

2.1.1 Convert into gray scale

First if the original image is RGB image then the pre-processing stage starts with converting the original RGB image into a grayscale image.

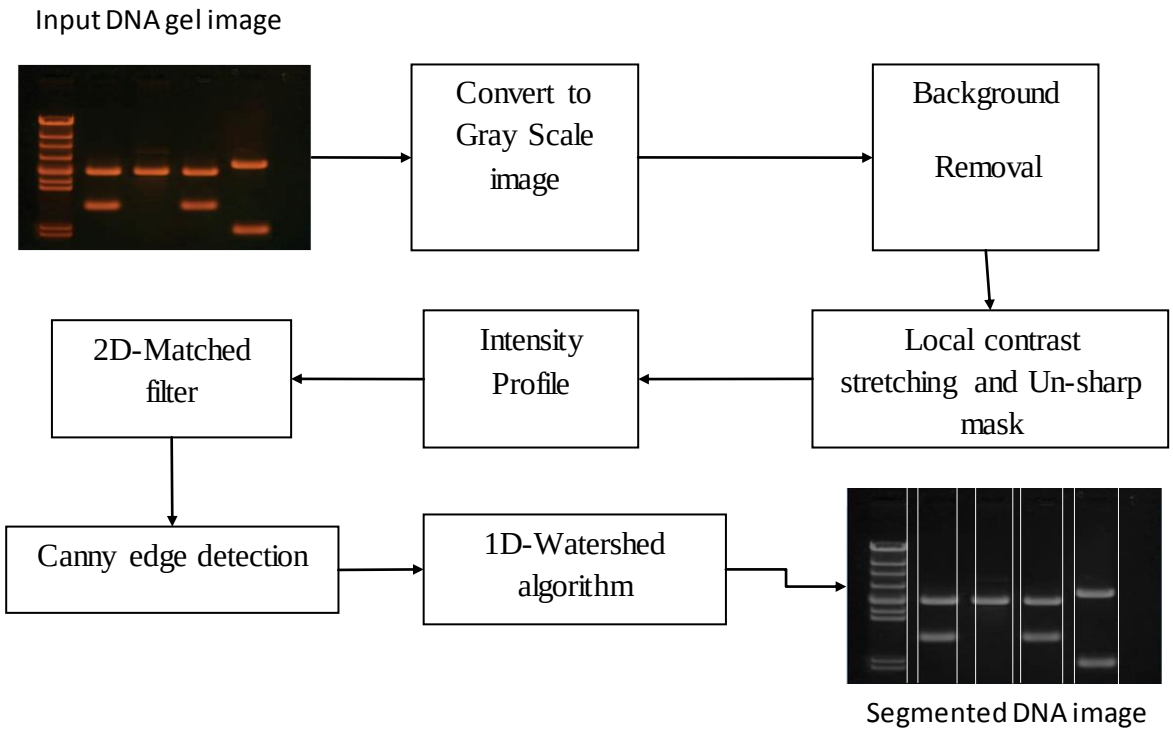


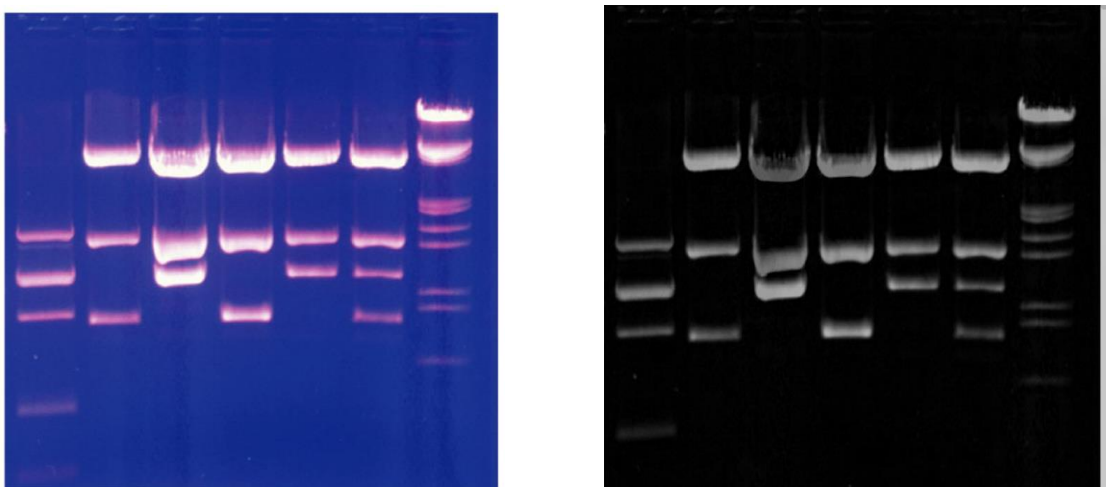
Figure1: Block diagram for the proposed scheme

2.1.2 Background removal

In the preprocessing stage we have to remove the background, these background pixels generally have a lower intensity than the pixels in the bands. The eroded image shows major variation in the background. Eroded image is then subtracted from the original image. This filter reduces the background variation considerably. Figure 2, shows the original image and the image after background subtraction where the bands are more clearly visible on it.
Next step is to further improve the contrast between DNA bands and the background. We have proposed local contrast enhancement filter for this purpose. This filter suppresses most of the noisy gray variation in the background while enhancing the pixels it considers as belonging to the bands in the image.

2.1.3 Image enhancement

Image enhancement is the improvement of the visibility of an image or any portion of the degraded image, it sharpens image features such as edges, boundaries, or contrast to give an image that is more suitable than the original image. One of the most common degradations of the gel images, is its poor contrast. We can define the contrast of an image as the difference between its highest and lowest intensity values. Usually a histogram equalization technique used to solve the poor contrast in the degraded image, but here we used the local contrast stretching technique because it improves the brightness differences uniform across the dynamic range of the image.



(a) (b)

Figure: 2 DNA gel electrophoresis image (a) Original DNA image (b) DNA gel electrophoresis image after the background removal

2.1.4 Local contrast stretching (LCS)

Local contrast stretching (LCS) is an enhancement method performed on an image for locally adjusting each picture element value to improve the visualization of structures in both darkest and lightest portions of the image at the same time [9]. LCS is performed by sliding windows (called the KERNEL) across the image and adjusting the center element. The contrast enhancement filter finds the average brightness in the larger surrounding region and subtracts it from the average brightness in the interior region.

$$D(x,y) = f((x,y) - min)/(max - min)) * N \quad (1)$$

Where N is the number of intensity levels, "min" and "max" are the minimum intensity value and the maximum intensity value in the input image. For example, normally in the gray-level standard, the lowest possible intensity is 0, and the highest intensity value is 255. Thus N is equal to 255. After enhancing the DNA gel image by using LCS, apply the Un-sharp masking that yields to increase either sharpness or local contrast.

2.1.5 Un-sharp Mask

An "un-sharp mask" is actually used to sharpen an image, contrary to what its name might lead you to believe. Sharpening can help you emphasize image details. The un-sharp mask filter is an extremely versatile sharpening tool that improves the definition of fine detail by removing low-frequency spatial information from the original image.

$$F(x,y) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \quad (2)$$

In general, increasing the size of the kernel mask causes the Gaussian filter to remove a greater number of spatial frequencies from the un-sharp mask image. The un-sharp mask is then subtracted from the original image according to the equation:

$$F(x,y) = \frac{c}{2c-1} I(x,y) - \frac{(1-c)}{2c-1} U(x,y) \qquad (3)$$

Where F (x, y) represents the brightness value of a pixel at the coordinate (x, y) in the filtered image, and I (x, y) and U (x, y) represent the brightness values of the corresponding pixels in the original and un-sharp mask (blurred) images, respectively. The constant "c" controls the relative weightings of the original and blurred images in the difference equation. Here the optimal value of c is between (1 to 5/9 (0.556))

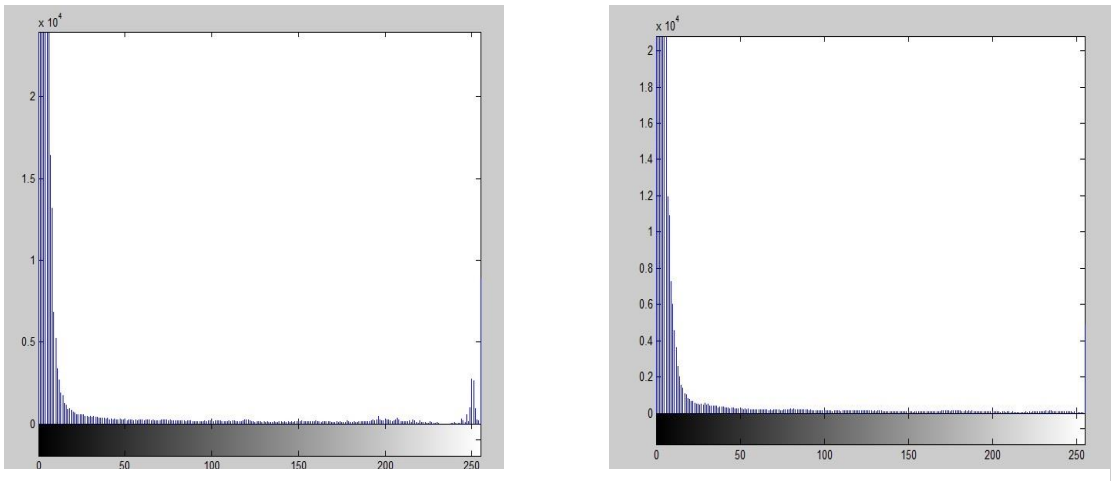
One of the primary advantages of the un-sharp mask filter over other sharpening filters is the flexibility of control, because a majority of the other filters do not provide any user-adjustable parameters. By the un-sharp mask filter, the preprocessing stage is finished. Figure 3 (a) and 3 (b) shows the histogram of the original image and the histogram of the image after the preprocessing respectively, and Figure 4 (a) and (b) shows the image before and after the preprocessing respectively.

2.2 Lane detection

Many applications in signal processing and image processing are involved with determining and detecting the presence and location of a target signal within some other signal. We first use the intensity profile to determine the lane location, then the watershed segmentation algorithm. Direct application of the watershed segmentation algorithm leads to over-segmentation due to noise and other local irregularities of the gradient. A practical solution to this problem is to limit the number of allowable regions by incorporating a preprocessing stage designed to bring additional knowledge into the segmentation procedure. This preprocessing stage is the matched filter that enhances the bands shape.

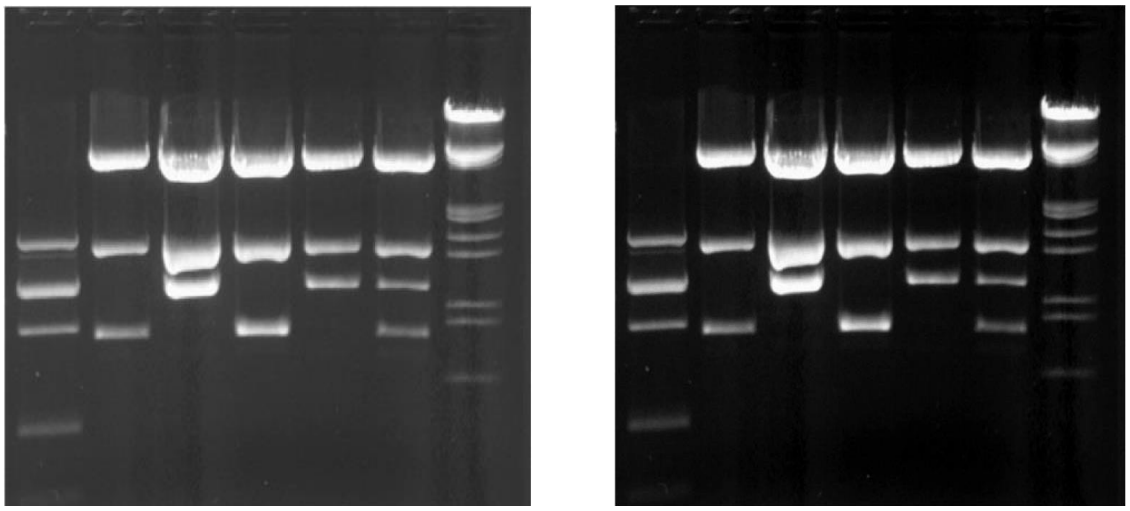
2.2.1 Intensity profile

In order to detect the position of the lanes, we have to find the image intensity profile of the image. The intensity profile obtained by the projection of column intensities in the image onto the horizontal axis (axis



(a) (b)

Figure: 3 Histogram of the DNA gel image (a) Histogram of the original DNA gel image (b) Histogram of the DNA gel image after the preprocessing



(a) (b)

Figure 4: DNA electrophoresis gel image (a) gray scale DNA gel image (b) DNA gel image after the preprocessing stage.

perpendicular to the lanes in the image). Then the intensity profile obtained by applying in (4)

$$p(x) = \frac{1}{N}\sum_{y=1}^{N} I(x,y) \qquad x = 1,..,M. \qquad (4)$$

Where the intensity profile P (x) of image I (x, y) of size M* N (where M is the number of rows and N is the number of columns in the image, this leads to identify the lane positions in a gel image as shown in figure 5 where the red dot represent the lane detection.

2.2.2 Matched filter

The matched filter technique is widely used in signal processing to increase the SNR. An introduction to the matched filter concept can be found in [10]. One requirement for using a matched filter is the need to construct the filter based on the knowledge of the intensity profile of the image. The matched filter is an optimal linear filter under the assumptions of Gaussian distribution of the signal to be detected.

According to the optimality of the matched filter with the assumption of the Gaussian noise, we used it to enhance the blurred bands. The matched filter can be approximated by the Gaussian curve as in (5)

$$D(y) = e^{\frac{-y^2}{2\sigma^2}} \ , \ -\infty \leq y \leq \infty \qquad (5)$$

Because of the above mentioned assumption, instead of matching a single intensity profile, a significant improvement can be achieved by matching a number of cross sections of identical profiles simultaneously. Thus, a 2-D matched filter was applied due to the bands which have concaved shape that made it rectangular shaped (two dimensional), and not horizontal line segments, its mathematical expression is

$$D(,,\mathrm{y}) = e^{\frac{-y^2}{2\sigma^2}} \ -\frac{d_y}{2} \leq y \leq \frac{d_y}{2} \ , -\frac{d_x}{2} \leq x \leq \frac{d_x}{2} \qquad (6)$$

Where d_x is the width and the height is d_y, Those two parameters must be determined for the matched filter. Since the bands are not perfectly straight lines, d_x width should not be as the length of the bands. In this experiment, $d_x = 5$ is an appropriate value for most of the cases. The height d_y depends on the variance σ^2 in (7). σ is varying depending on the location of the band, because the bands closer to the top are wider than those closer to the bottom of the image. We set σ as a linear function of y, as follows

$$\sigma = 1 + c.\left(\frac{y}{N}\right) \qquad (7)$$

Where y is the distance between the band and the bottom side of the image. The Gaussian distribution quickly drops to zero when σ is small, and so d_y should be small. Conversely, for a large σ, the Gaussian distribution slowly becomes zero and so d_y should be large. To determine d_y from a given σ we used the method on [11]. The height is between $-(4\sigma + 3)/2 \leq y \leq (4\sigma + 3)/2$. Then the matched filter equation is

$$D(x,y) = e^{\frac{-y^2}{2\sigma^2}}, -\frac{d_y}{2} \leq y \leq \frac{d_y}{2} \ , -\frac{d_x}{2} \leq y \leq \frac{d_x}{2} \ , \sigma = c.\left(\frac{y}{N}\right) \qquad (8)$$

Where $d_y = 4\ (\sigma + 3)$ and $d_x = 5$. Applying the matched filter to the DNA gel images makes the bands enhanced. The result after applying the matched filter is shown in Figure 6. The average intensity profile for each lane before and after applying the matched filter is shown in Figure 7 (a) and (b) respectively.

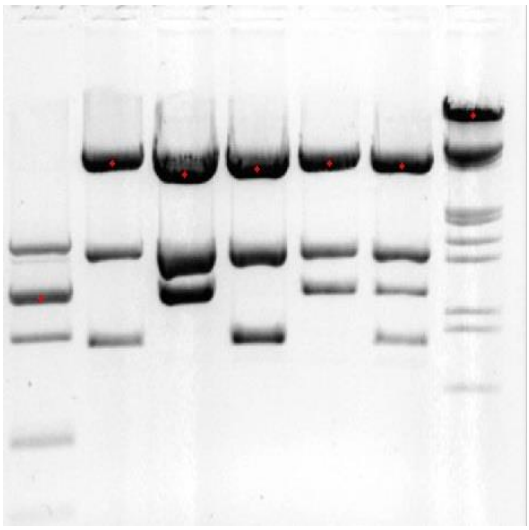


Figure 5: lanes detection in the DNA gel image

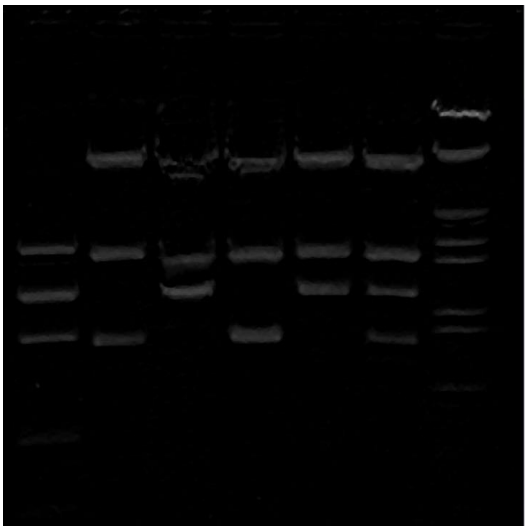
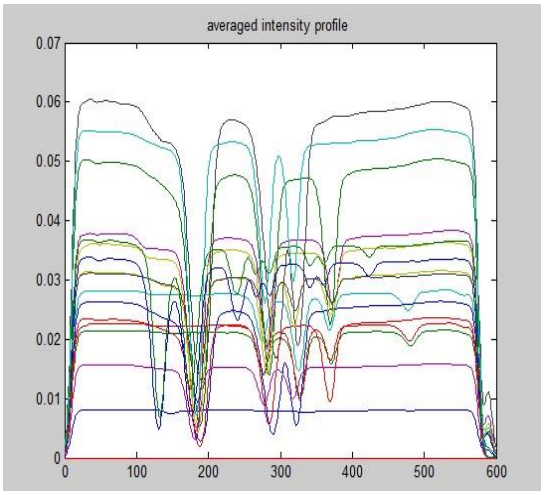
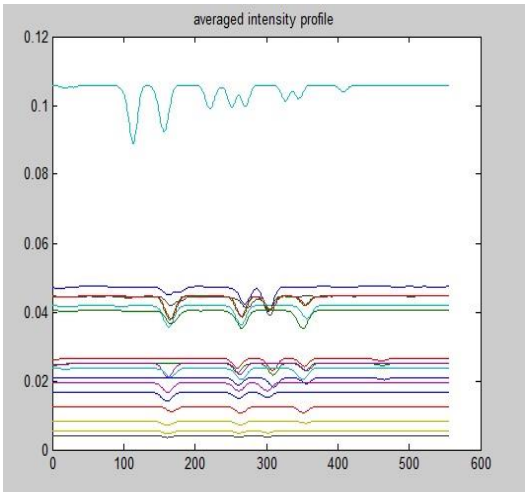


Figure 6: The matched filter of the DNA gel image



(a)



(b)

Figure 7: Average intensity profile of the DNA gel image (a) Before applying the matched filter (b) After applying the matched filter

2.3 Lanes segmentation

Lane and band segmentation is difficult due to the quality of the GE images. It is important to identify the lanes before segmentation. Image Segmentation is the process of partitioning a digital image into multiple regions or sets of pixels [12] [13].

2.3.1 Watershed Segmentation Algorithm

The watershed algorithm is a well established morphological segmentation tool. It commonly segments an image into a set of non overlapping regions. It also has the advantage

of a region growing algorithm, the regions are spatially consistent, with boundaries forming a closed, connected set, and it also makes use of edge information, as captured by the gradient surface. First we have to find the gradient image by using a canny edge detector as shown in Figure 8.

The point on the centerline of a band is the peak of intensity profile of the lanes. Thus, the center of a band can be found by determining the local maxima (peaks) on the profile. Since the peaks do not have the same height therefore, the intensity threshold is not applicable here. Watershed algorithm [14] - [16] is usually employed for local maxima determination and image segmentation; we used it here to find the peaks instead of the intensity threshold. It aims to find the peaks in the image gradient called watersheds and identifying them as the image contours. Therefore a 1-D watershed algorithm is applied to determine the peaks of all the vertical scan lines in the image. Where these peaks represent the center of the bands, Figure 9 shows the segmented DNA gel image.

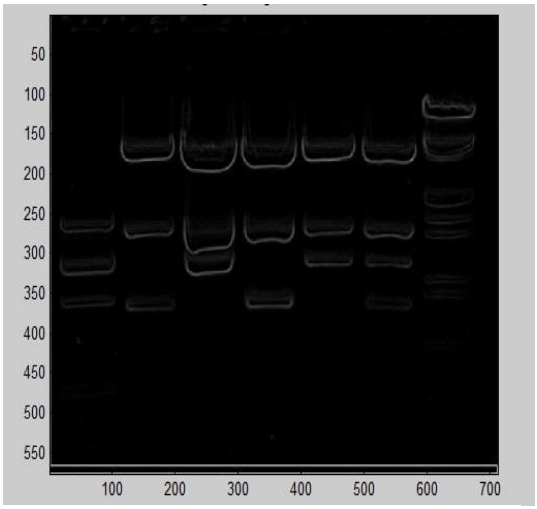


Figure 8: the gradient image by using canny edge detector.

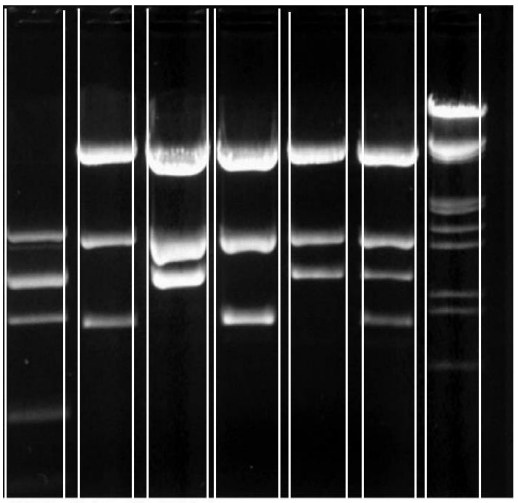


Figure 9: the segmented DNA gel image using the watershed algorithm.

3. Experimental Results and Evaluation

In order to test the accuracy of the proposed scheme we have tested it on a wide variety of the DNA gel electrophoresis images. Experimental results show that the proposed scheme is able to detect all the lanes successfully and up to 99.5% accuracy for the segmentation of lanes in good quality images.

3.1 Data Set

In order to test the accuracy and efficiency of the proposed scheme, we have used 20 test images of both single- locus and multi-locus DNA, with various qualities, resolution and with different numbers of lanes.

3.2 Evaluations

There are many segmentation evaluation methods. However, segmentation evaluation is still an open topic [17], [18]. To evaluate the performance of the proposed scheme, objective evaluation tools are used, such as mean square error (MSE) and peak signal to noise ratio (PSNR) are used. Equations (9) and (10) below represent the MSE and PSNR respectively.

$$MSE = \frac{1}{MN}\sum_{J=1}^{M}\sum_{k=1}^{N}\left(x_{J,k} - x'_{J,k}\right)^2 \qquad (9)$$

$$PSNR = 10\,log\frac{(2^n-1)^2}{MSE} = 10\,log\frac{255^2}{MSE} \qquad (10)$$

The results of Equations (9) and (10) for the proposed scheme compared with iterative moving average IMA algorithm (that depends on the image quality and is also sensitive to the distance between two adjacent lanes in a gel image) are shown in Figure 10 (a) and (b) respectively.

4. Conclusion and Future Work

In this paper, the gel electrophoresis analysis scheme provides a fast and easy way to detect and segment the lanes. We have presented a scheme, which can be independently used for gel electrophoresis image analysis or as a tool for cross checking the manually called results. In theory, due to large amounts of data to be analyzed, high throughput bio-informatics procedures should be fully automatic. It is too serious to design fully automatic systems when the quality of data input varies and the human factor is involved in gel preparation and image acquisition.

In the future we intend to develop and improve a method to define a classification system, to develop an automatic method for the analysis of DNA fingerprint gel images that extracts a meaningful set of features from each lane. This will allow the automatic recognition and classification of patterns, in normal or pathological cases.

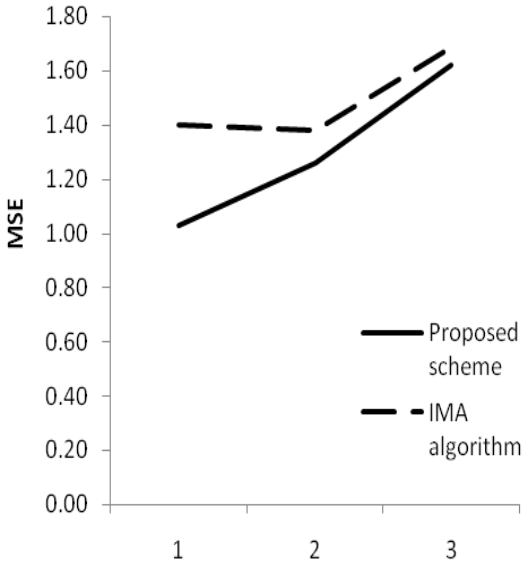


Figure 10: MSE for the proposed scheme and the IMA algorithm.

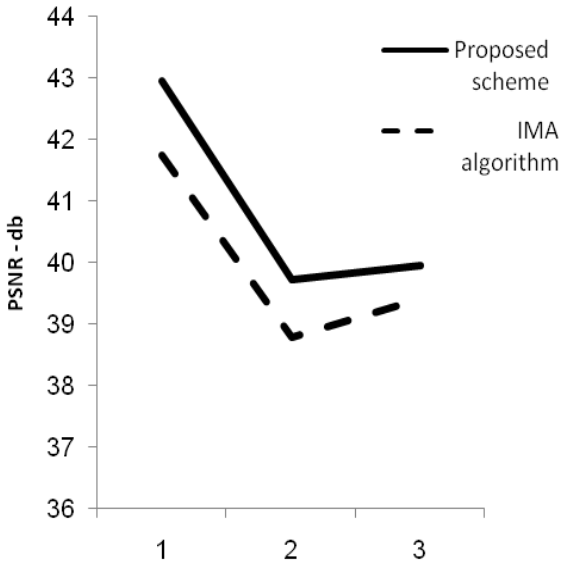


Figure 11: PSNR for the proposed scheme and the IMA algorithm

References

1. J. Sambrook, and D.W. Russell, "Molecular Cloning: A laboratory Manual. Cold Spring Harbor, Cold Spring Harbor Laboratory ", NY Press, 3rd Ed, 2001.
2. JK. Elder, and EM. Southern "Computer -aided analysis of one dimensional restriction fragment gels, in Nucleic acid and protein sequence analysis", Ed. M. J. Bishop and C. J. Rawlings, IRL oxford Press, 165–172, 1987.
3. Kaabouch, Naima, Richard R. Schultz, and Barry Milavetz. "An analysis system for DNA gel electrophoresis images based on automatic thresholding an enhancement." In Electro/Information Technology, IEEE International Conference on, pp. 26-31. IEEE, 2007.
4. A. Akbari, and F. Albregtsen. "Evaluation of noise in DNA fingerprint images produced by hybridization techniques." In Proceedings of the 6th Nordic Signal Processing Symposium-NORSIG, vol. 2004. 2004.
5. L. Chih-Yang, Y. Ching, and Y. Liang Yang. "An automatic method to compare the lanes in gel electrophoresis images." Information Technology in Biomedicine, IEEE Transactions on 11.2, pp. 179-189, 2007.
6. C. Wei-Zen, K. Yen, C. Yang Lin, Y. Ching, and Y. Yang. "Comparing lanes in the pulsed-field gel electrophoresis (PFGE) images." In Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE, vol. 3, pp. 2911-2913. IEEE, 2001

7. A. Akbari, F. Albregtsen, and S. Jakobsen. "Automatic lane detection and separation in one dimensional gel images using continuous wavelet transform." Analytical Methods 2, no. 9, pp. 1360-1371, 2010.
8. Rapley, Ralph, and John M. Walker, eds. Molecular Biomethods Hand Book. No. 1646. Humana Press Inc, 1998.
9. Gonzales, C. Rafael and R. E. Woods. "Digital image processing, 1993."
10. G. Turin, "An introduction to matched filters." Information Theory, IRE Transactions on 6, no. 3, pp.311-329, 1960.
11. G. Lohmann, Volumetric image analysis. Wiley, 1998.
12. M. Abdulghafour, "Image segmentation using Fuzzy logic and genetic algorithms." Journal of WSCG 11, no. 1 , 2003.
13. N. Senthilkumaran and R. Rajesh. "Edge detection techniques for image segmentation-a survey of soft computing approaches." International Journal of Recent Trends in Engineering 1, no. 2, pp. 250-254, 2009.
14. J. B. Roerdink, , and A. Meijster. "The watershed transform: Definitions, algorithms and parallelization strategies." Fundamenta Informaticae 41, no. 1, pp. 187-228, 2000.
15. L. Vincent, and P. Soille. "Watersheds in digital spaces: an efficient algorithm based on immersion simulations." IEEE transactions on pattern analysis and machine intelligence 13, no. 6, pp. 583-598, 1991.
16. S. Beucher, and C. Lantuéjoul. "Use of watersheds in contour detection." 1979.
17. S. Ruan, S. Lebonvallet, and T. Qiu. "Automatic image segmentation based on level set approach: application to brain tumor segmentation in MR images." 2009.
18. H. Zhang, J. E. Fritts, and S. A. Goldman. "Image segmentation evaluation: A survey of unsupervised methods." Computer Vision and Image Understanding 110, no. 2, pp. 260-280, 2008.