ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 9, No. 4, 2014, *pp*. 243-251

Bands detection and Lanes segmentation in DNA Fingerprint images

I. Ismail, Gh. S. Eltaweel, H. Nassar
Department of Computer Science
Faculty of Computers & Informatics
Suez Canal University, Ismailia 51422, Egypt
Is-raa@hotmail.com, Ghada_eltawel@ci.suez.edu.eg,
(*Received February 10, 2012, accepted July 23, 2014*)

Abstract. Gel electrophoresis (GE) is a widely used technique to separate DNA sequence according to their size and weight, GE results are presented using images. In this paper, we present a scheme that aims to detect and segment the lanes in DNA gel images without any human interventions. We have successfully implemented an image processing scheme to automatically detect and segment the lanes in DNA gel images, applying this scheme to several DNA fingerprint gel images ,all the lanes are successfully segmented. Also we have obtained up to 99.5% accuracy for the segmentation of lanes in good quality images. The proposed scheme is compared with other techniques and the comparison shows that it has a minimum error rate.

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

1. Introduction

Gel Electrophoresis (GE) is a method used in clinical chemistry to separate proteins by charge and/or size also in \ble tool in many applications such as forensic studies, paternity analysis, protein profile comparison, gene identification, isolation, purification and population genetic analysis.

GE technique generates images that called DNA fingerprinting images which are an efficient and highly accurate means of identities and relationships. DNA fingerprinting images consist of several vertical *lanes*, each lane corresponding to one sample. Each lane contains a number of horizontal *bands*. Each band represents a part of the sample. The positions of the horizontal bands in the lane represent the molecular weights of that part of the sample. Two samples are considered to be the same if their lanes have the same pattern, Fig. 1 shows an example of a DNA fingerprint gel electrophoresis image.

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 equispaced 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 main steps: automatic thresholding, shifting, filtering and data processing. They use the automatic thresholding to equalize the grey values of the gel electrophoresis image background.

Akbari A. et al [4] presented an effective noise filtering technique for DNA gel images. Lin, C.Y et al [5] designed a computer 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 that 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 detect and segment DNA fingerprint gel images using image processing techniques to automate routine analysis process of GE DNA images to help identifying humans.

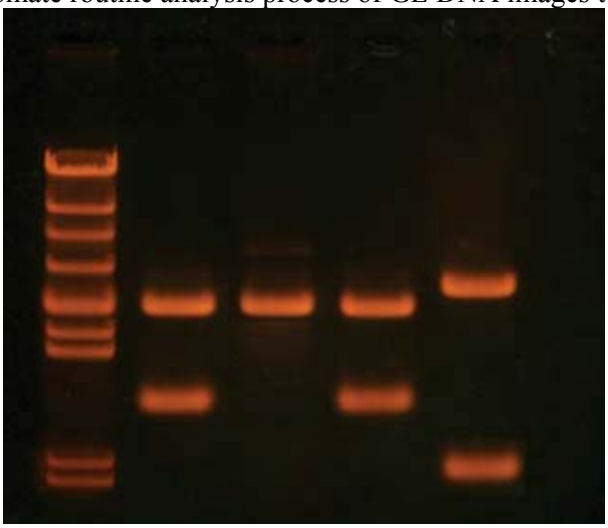


Fig.1 DNA fingerprint gel electrophoresis image.

Automatic tools speed up the routine of the biological processes. Much repetitive work in molecular biology can be in a computerized analysis form that is reproducible and avoids various forms of human error. Automatic techniques with an interactive check on the results speed up the analysis and reduce the error, and that's the motivation of this scheme.

2. Materials and Methods

The proposed scheme consists of two stages, the pre-processing stage starts by converting the image into a gray scale image, then image enhancement and background removal, The main purpose of the pre-processing stage is to enhance the image and make it sharper, The lanes detection and segmentation stage, starts by using 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. Fig. 2 shows the block diagram for the proposed scheme.

2.1 Preprocessing

The pre-processing stage starts with converting the original RGB image into a grey scale image, as shown in fig.3 (a), and then image enhancement process is applied to improve the visual appearance of an image and converts the image to a form better suited for analysis by a human or a machine. Using un-sharp / local contrast stretching [9], because most of the un-sharp masking methods are not effective for low-contrast images. Therefore, after converting the image into gray scale, the contrasts of these images are stretched.

The main objective of using local contras stretching is to highlight faint bands that can be washed out because of their low gray levels. 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. LCS is performed by sliding windows (called the kernel) across the image and adjusting the center element using the formula

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

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 because these are both forms of increasing differences between values and increasing slope

sharpness referring to very small scale (high frequency) differences, and contrast referring to larger scale (low frequency) differences.

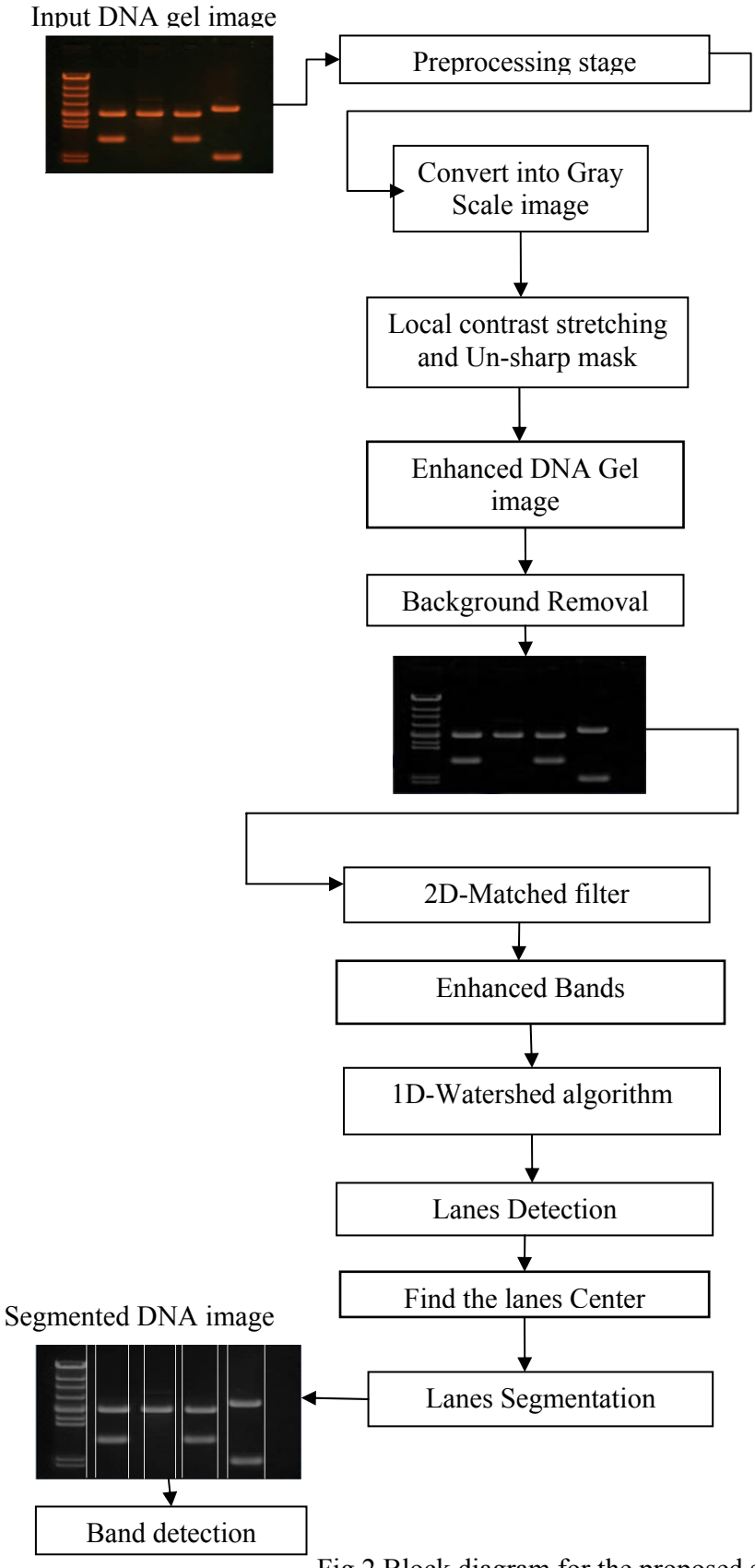


Fig.2 Block diagram for the proposed scheme

The second step in the preprocessing stage is to set the intensity of the pixels not in the bands to zero. These background pixels generally have a lower intensity than the pixels in the bands. In Fig. 4(a) the histogram approximately consists of two normal distributions. A threshold is set to be the closest gray level corresponding to the minimum probability between the maxima of two normal distributions, which results in minimum error segmentation. Such optimal threshold was solved by using the method proposed in [10]. The gel image after the preprocessing stage is shown in Fig.3 (b). And the histogram after the preprocessing is shown in Fig. 4(b)

2.2Lanes detection and Segmentation

Lane and band segmentation is difficult due to the quality of the GE images. It is important to identify the lanes before segmentation. Segmenting bands and lanes is designed based on these properties of bands and lanes.

1. Bands which are closer to the top are wider than those closer to the bottom of the image.
2. Bands in the same image but in different lanes may have different shapes.
3. Bands shapes are similar in the same lane and are concaved downward.
4. Band could break into several fragments due to noise.

To enhance the blurred bands, the matched filter technique [11]–[14] is employed.

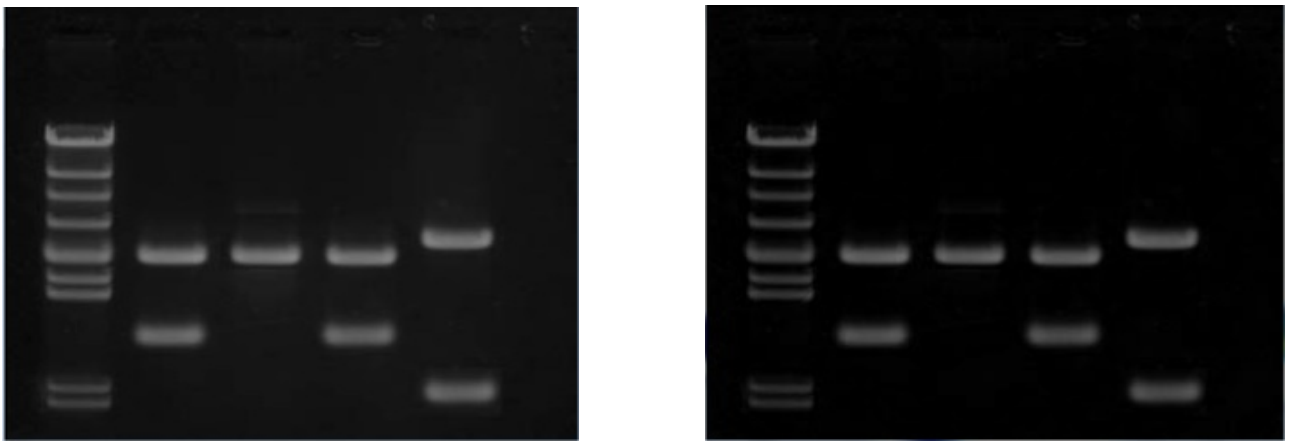
Matched filter design is based on the shape and intensity distribution of the bands; it yields to a high signal-to-noise ratio (SNR). A large response can be obtained if the matched filter is applied to a place Y where there is a band. A band profile is bell-shaped and can be approximated using a Gaussian distribution in the y direction, as follows

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

We have to apply a 2-D matched filter due to the bands concaved shape that made it rectangular-shaped (two dimensional), and not horizontal line segments, the 2-Dmatched filter equation is needed as in (3)

$$D(\mathrm{x},\mathrm{y}) = e^{\frac{-y^2}{2\sigma^2}} - \frac{d_y}{2} \leq y \leq \frac{d_y}{2}, -\frac{d_x}{2} \leq x \leq \frac{d_x}{2} \tag{3}$$

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 (4). 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.



(a) (b)

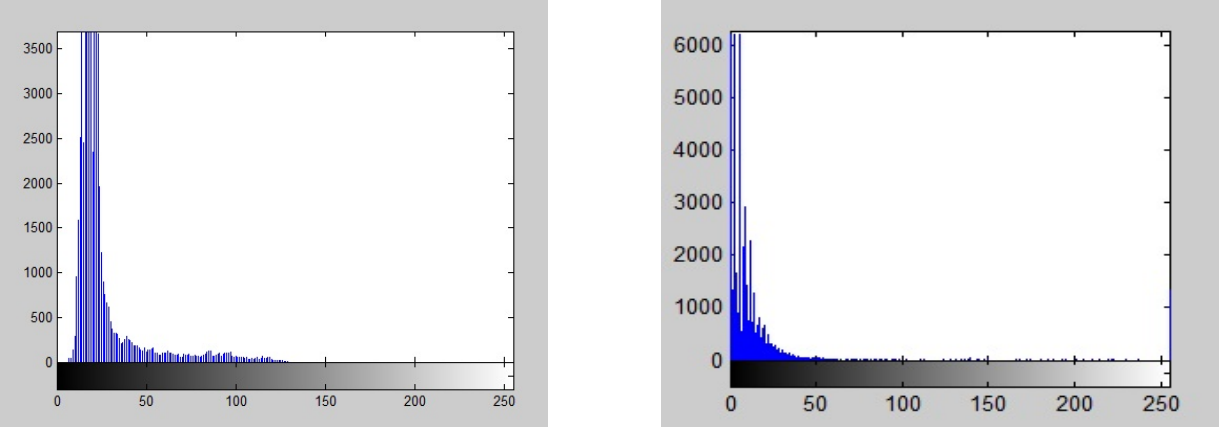
Fig.3 shows the DNA gel image (a) before preprocessing, (b) after the preprocessing.

We set σ as a linear function of y, as follows

$$\sigma = 1 + c.\left(\frac{y}{N}\right) \tag{4}$$

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 [15]. The height is between $-(4\sigma + 3)/2 \leq y \leq (4\sigma + 3)/2$. Then the matched filter equation is



(a) (b)

Fig.4 Histogram of the DNA gel image (a) Original image, (b) After the preprocessing.

$$D(\mathrm{x},\mathrm{y}) = e^{\frac{-y^2}{2\sigma^2}} \quad -\frac{d_y}{2} \leq y \leq \frac{d_y}{2} \;\; , \;\; -\frac{d_x}{2} \leq x \leq \frac{d_x}{2}, \;\; \sigma = c.\left(\frac{y}{N}\right) \tag{5}$$

Where d_y= 4(σ + 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 Fig.5. The average intensity profile for each lane before and after applying the matched filter is shown in Fig.6 (a) and Fig.6 (b) respectively.

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.

The watershed algorithm that also called "Catch basin algorithm" [16]-[18] that was introduced by Vincent and Soille is usually employed for local maxima determination and image segmentation in image processing, we used it here to find the peaks instead of the intensity threshold.

It produces a more stable segmentation of objects including continuous segmentation boundaries by a concept of producing catchment basin (watersheds) and watershed line (divide lines or dam boundaries). It is a well established morphological segmentation tool, that segment an image into a set of non overlapping regions. It also has the advantage of a region growing algorithm. It aims to find the peaks in the image gradient called watersheds and identifying them as the image contours. Therefore 1-D watershed algorithm is applied to determine the peaks of all the vertical scan lines in the image.

These peaks represent the center of the bands, which tend to form a connected component. Most of the connected components are found at the center of the bands while some are just noise. We can remove this noise by applying a size filter with a proper threshold. The lane detection step is shown in Fig.7 where the red dot indicates that it is a lane and the final result of the lanes detection and segmentation stage is shown in Fig. 8, where the white lines along the image represent the lanes segmentation of the DNA gel image.

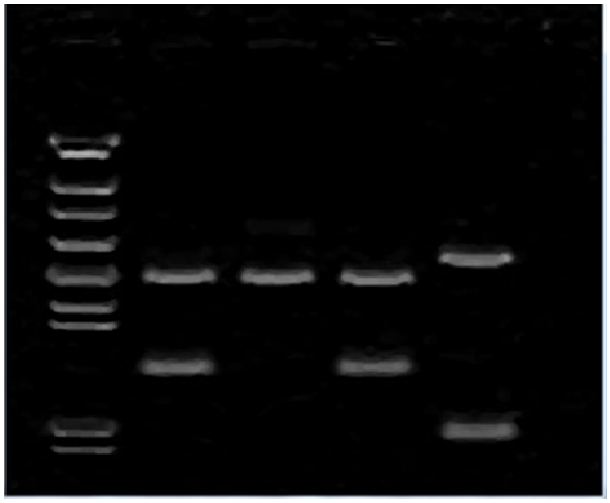
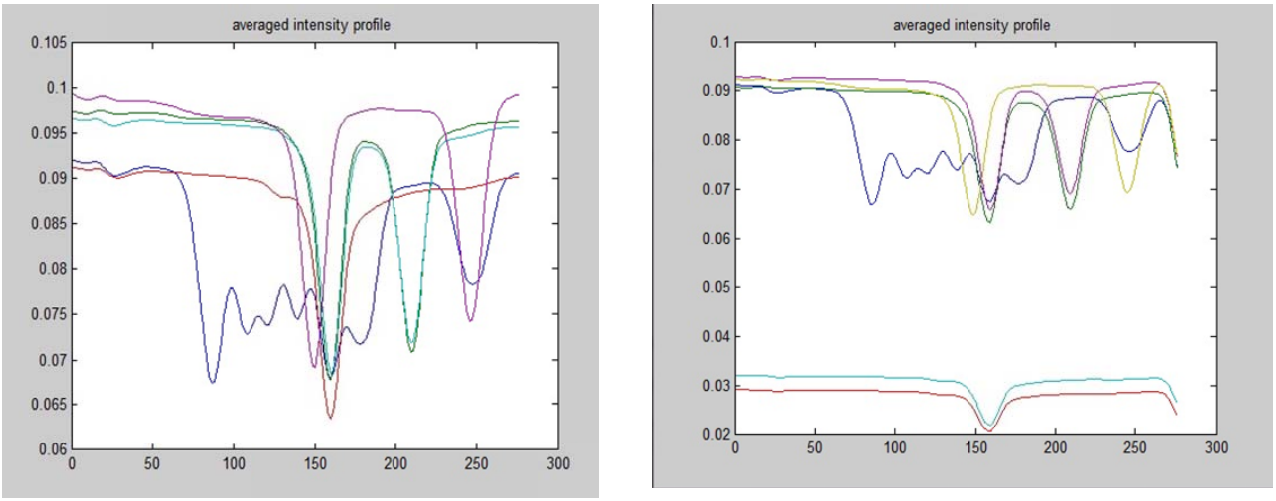


Fig.5 The DNA gel image after applying the matched filter.



(a) (b)

Fig.6 Average intensity profile of the DNA gel image (a) Before applying the matched filter (b) After applying the matched filter.

2.3 Bands Detection

Ideally, if a DNA band appears in the image, the row that represents the band's centerline should have a local minimum on the intensity profiles on each column in the image. But due to noise or distortions some of these local minima are either displaced or vanished. Based on the assumption that the centerline on a DNA band represents a local minimum on the intensity profile on each column in the image, we will be able to use the average intensity profile obtained from all columns in the image to isolate the DNA bands. The intensity profile of one electrophoresis band can be mathematically modeled by means of a Gaussian function as in (6)

$$D(x) = A\, e^{\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2} \tag{6}$$

Where μ represents the central band location, A the intensity level or amplitude on μ, and σdefines the width of the i^{th} band. While the shape of DNA bands is rectangular, then the average signal is first smoothed by a 1-D mean filter with a window size 3*3 to remove noise with small amplitude. Then the position for all local minima on the signal is computed. At this level we assume that each valley on the signal represents a DNA band. The information of these valleys is used to compute the boundary positions for the corresponding DNA band. The equivalent width of DNA fingerprint bands can be found in [19], the detection of the bands is shown in fig. 9.

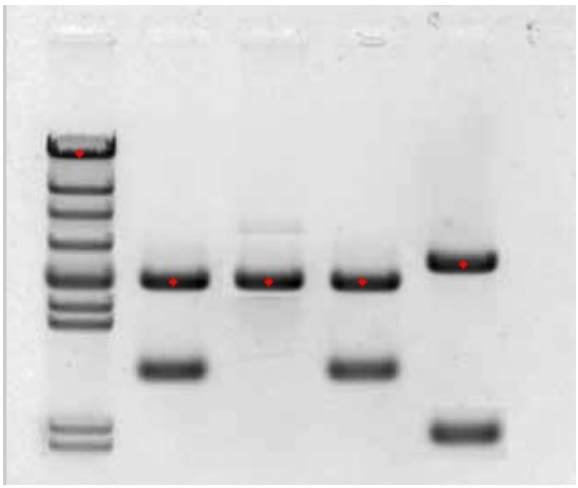


Fig. 7 the lanes detection

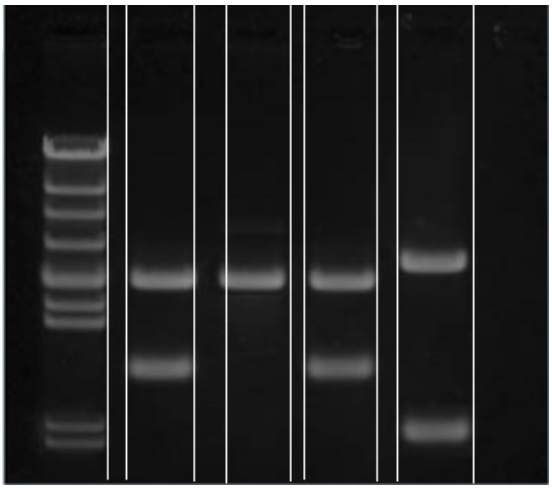


Fig. 8 The lane segmentation result

3. Experimental Results and Evaluation

Particle swarm optimization (PSO) as a parallel optimization algorithm can be used to solve a large number of complex and non-linear problems, and has been widely used to science and engineering for example function optimization, pattern classification and resource allocation fields [8]. PSO is most frequently applied to solve continuous optimization problems. At present, how to apply PSO to discrete optimization problems, especially combinatorial optimization problems, is an important research direction. In recent years, many researchers have proposed improved PSO for example in [12] for combinatorial optimization problems and obtained many good optimization solutions. Therefore, we also use PSO to find optimal quality control lines (local constraints).

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.

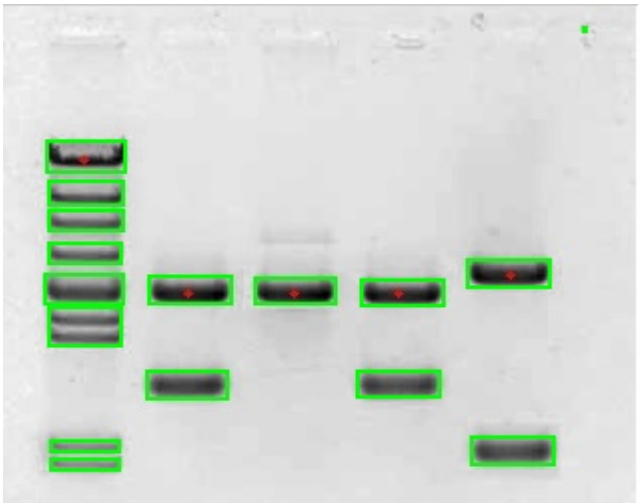


Fig.: 9, Bands detection of the GE DNA finger print image

3.1 Data Set

We have used 20 test images of both single- locus and multi-locus DNA, with various qualities, resolution between 150 dpi - 300 dpi and with different numbers of lanes.

3.2 Evaluations

There are many segmentation evaluation methods. However, segmentation evaluation is still an open topic [20], [21]. 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) [22] are used. Equations (6) and (7) below represent the MSE and PSNR respectively.

$$MSE = \frac{1}{MN} \sum_{J=1}^{M} \sum_{k=1}^{N} \left(x_{J,k} - x'_{J,k}\right)^2 \quad (6)$$

$$PSNR = 10\ log \frac{(2^n-1)^2}{MSE} = 10\ log \frac{255^2}{MSE} \quad (7)$$

The results of Equations (6) and (7) 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) and the Hough algorithm are shown in Fig.10 and Fig.11 respectively. In order to evaluate the lane detection method, the results for the image data set are expressed in terms of the confusion matrix presented in table I.

3.3 Comparison with other Schemes

When we compare the human expert and the proposed scheme, the proposed scheme is evidently much more sensitive. The comparison of the lane segmentation obtained by the proposed scheme and the human expert results are shown in table II.

Table I: The performance of the lane detection method expressed in terms of confusion matrix

	Detected	**Not detected**
True lanes	0.99	0.01
False lanes	0.02	0.98

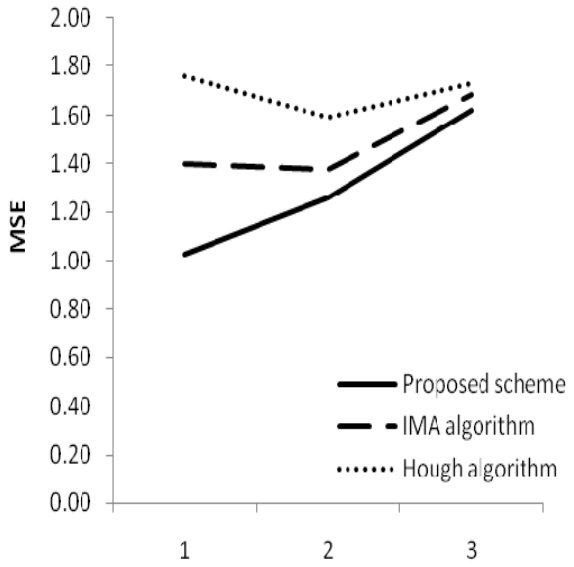


Fig.10 MSE for the proposed scheme, Hough algorithm and IMA algorithm.

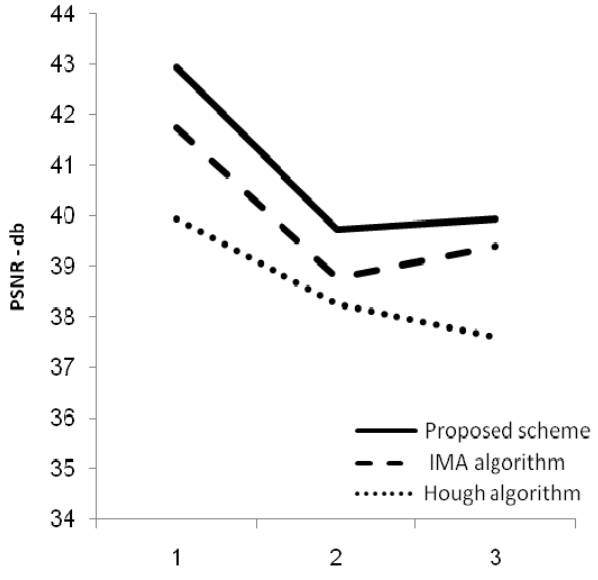


Fig.11 PSNR for the proposed scheme, Hough algorithm and IMA algorithm

Table II: Comparison of lanes Segmentation results obtained by the proposed scheme and a human expert

	false positive	**false negative**	**total error**	**error rate**
proposed scheme	9	9	18	2.4
experienced human expert	3	19	22	2.8

4. CONCLUSIONS

In this paper, an accurate lanes segmentation scheme for GE images is presented. This scheme detects and segments the lanes in gel electrophoresis images. There are a number of image processing steps that should be applied based on the performance of segmenting the lanes, tested this scheme on 20 GE image with various qualities, resolutions and different numbers of lanes, it shows that the proposed scheme has a higher performance with respect to high peak signal to noise ratio and minimum error rate compared with the Hough transformation. We expect that this scheme will help biologists save a great effort comparing GE images.

5. 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] J. K., Elder, and E.M., 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 Press,* 165–172 (1987).
[3] N. Kaabouch "An Analysis System for DNA Gel Electrophoresis Images based on Automatic Thresholding and Enhancement". *In IEEE EIT.* (2007).
[4] Akbari and F. Albregtsen "Evaluation of noise in DNA fingerprint images produced by hybridization techniques". *In Proc 6th – NORSIG*, (2004).
[5] C.Y. Lin, Y.T. Ching, and Y.L. Yang, "An Automatic Method to Compare the Lanes in Gel Electrophoresis (GE) Images". *IEEE Trans*, 11(2), pp. 179 - 189 (2007).
[6] W. Z. K. S. Yen, C.Y. Lin, Y. T. Ching, and Y. L. Yang "Comparing lanes in the pulsed-field gel electrophoresis (PFGE) images". *In Proc.23rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc*. 2911–2913 (2001).
[7] Akbari, F. Albregtsen and K. S. Jakobsen "Automatic lane detection and separation in one dimensional gel images using continuous wavelet transform". *Analytical Methods,* 2 (9), 1360 -1371 (2010).
[8] Smith, R. Duncan, "Gel Electrophoresis of DNA", *Molecular Biomethods Handbook* pp: 17-33 (1998). R. C. Gonzalez, and R. E. Woods, Digital Image Processing. New York: Addison-Wesley. 1992.
[9] C. A. Glasbey, "An analysis of histogram-based thresholding algorithms". *CVGIP: Graph. Models Image Process*, 55, 532–537(1993).
[10] L. A. Wainstein, V. D. Zubakov, and D. Hildereth "Extraction of Signals from Noise". *Englewood Cliffs, NJ: Prentice-Hall* (1962).
[11] G. L. Turin, "An introduction to matched filter". IRE Trans. Inf. Theor, 1960 6, pp. 311–329
[12] D. Middleton "On new classes of matched filters and generalizations of the matched filter concept". *IRE Trans. Inf. Theor* , 6, 349–360(1960).
[13] P. V. Villeneuve, H. A. Fry, J. Theiler, B. W. Smith and A. D. Stocker "Improved matched-filter detection techniques,". *Proc. SPIE*, 3753, 278–285(1999).
[14] G. Lohman, "Volumetric Image Analysis". New York: Wiley. 1998.
[15] J. B. T. M. Roerdink, and A. Meijster, "The watershed transform: Definitions, algorithms and parallelization strategies". *In Fundamenta Informatics. Groningen, the Netherlands: Inst. Math. Comput. Sci. Univ.,* 187–228 (2000).
[16] L. Vincent, and P. Soille, "Watersheds in digital space: An efficient algorithm based on immersion simulations". *IEEE Trans, Pattern Anal. Mach. Intell*, 13(6), 583–598 (1991).
[17] S. Beucher, and C. Lantu "Use of watersheds in contour detection". In Proc. International Workshop on Image processing, Real-Time Edge and Motion Detection/Estimation, Rennes, (1979).
[18] Akbari A.,"Lane detection and separation in DNA fingerprint gel images",NorSig, pp. 6-11, (2001).
[19] L. Xiaobing, "Automatic image segmentation based on level set approach: application to brain tumor segmentation in MR images", *Dalian University of Technology* (2009).
[20] H. Zhang, J.E. Fritts, and S.A. Goldman "Image segmentation evaluation: a survey of unsupervised methods". *Computer Vision and Image Understanding,* 110 (2), 260-280 (2008).
[21] http://en.wikipedia.org/wiki/PSNR