A Fuzzy-Watershed Based Algorithm for Protein Spot Detection in 2DGE images

Shaheera Rashwan[†], **Talaat Faheem**[††], **Amany Sarhan**[†††] **and Bayumy A.B.Youssef**[††††],
[†, ††††] Informatics Research Institute, Mubarak City for Science and Technology, Borg ElArab, Alexandria,Egypt
[††, †††] Computer Science Department, Faculty of Engineering, Tanta University, Tanta,Egypt

Summary
An important issue in the analysis of two-dimensional electrophoresis images is the detection and quantification of protein spots. The main challenges in the segmentation of 2DGE images are to separate overlapping protein spots correctly and to find the abundance of weak protein spots. To enable comparison of protein patterns between different samples, it is necessary to match the patterns so that homologous spots are identified. In this paper we describe a new robust technique to segment and model the different spots present in the gels. The watershed segmentation algorithm is modified to handle the problem of over segmentation by initially partitioning the image to mosaic regions using the composition of fuzzy relations. The experimental results showed the effectiveness of the proposed algorithm to overcome the over segmentation problem associated with the available algorithms.

Keywords: *Protein Spot Detection, Watershed Segmentation, over-segmentation, Fuzzy Relations*

1. Introduction

Two-dimensional gel electrophoresis (2-D Gel) enables separation of mixtures of proteins due to differences in their isoelectric points (pI), in the first dimension, and subsequently by their molecular weight (MWt) in the second dimension as sketched in Fig. 1. Proteins are separated in two dimensions; horizontally by iso-electric point (pI) and vertically by molecular weight (MWt). No proteins are shown. The pI and MWt ranges are example values.

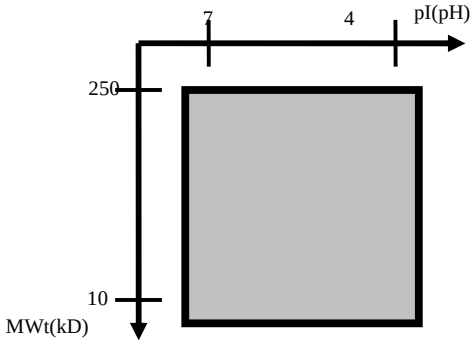

Fig. 1: Schematic two-dimensional electrophoresis gel

Other techniques for protein separation exist, but currently 2-D Gel provides the highest resolution allowing thousands of proteins to be separated.

The great advantage of this technique is that it enables, from very small amounts of material, the investigation of the protein expression for thousands of proteins simultaneously.

After protein separation, an image of the protein spot pattern is captured. Proper finding and quantitation of the protein spots in the images and subsequent correct matching of the protein spot patterns allows not only for the comparison of two or more samples but furthermore makes the creation of an image database possible[7,8].

Proteomics is an increasingly important part of cell biology and the efforts to understand the basic principles of life - how the living cell works.

In proteome analysis, gel electrophoresis is a technique to separate proteins in a biological sample on a gel. The resulting gel images are made by captured as a digital image of the gel. This image is then analyzed in order to quantitate the relative amount of each of the proteins in the sample in question or to compare the sample with other samples or a database.

The task of analyzing the images can be tedious and is subjective (dependent on the human operator) if performed manually.

The use of digital image analysis in the field of proteomics is primarily motivated by the need to improve speed and consistency in the analysis of two-dimensional electrophoresis gel (2-D Gel) images.

In this paper, the most important issues and challenges related to digital image analysis of the gel images will be addressed in this paper, namely the segmentation of the images.

We will present the application of the watershed algorithm to the two-dimensional electrophoresis gel (2-D Gel) images. However, such algorithm suffers from the over-segmentation problem. In order to overcome such problem, we propose the use of fuzzy notion to the original algorithm.

In [12], Hoang et al. presented a novel approach for protein spot detection, which is a marker-free Watershed that does not require specification of predefined markers for the process of finding watershed contour lines. This

Manuscript received May 5, 2010
Manuscript revised May 20, 2010

approach includes a selective nonlinear filter and pixel intensity distribution analysis for removing local minima which causes over-segmentation when applying watershed transform. It then superimposes those true minima over the reconstructed gradient image before applying Watershed transform for spot segmentation. The effectiveness of this marker-free approach was experimentally comparable with other methods.

In [13], Lin and kuo have developed an adaptive mechanism to adjust the level of detail and determine the threshold value of watershed. The over-segmentation draw back is overcome by applying directed graph version of watershed transform algorithm and morphological opening operation. Labeling and region growing techniques were adapted to extracted individual spots features.

In [14], the watershed algorithm was used for spots segmentation in 2DGE images. But the paper is more focused on using the diffusion principle in modeling the spots.

In [15], marker-based watershed segmentation methods were used to improve the segmentation of the protein spots from the varying background.

In our work, we will introduce the notion of fuzzy relations to handle the problem of over-segmentation often produced by the watershed algorithm

This paper is organized as follows: section 1 presents the introduction. Section 2 summarizes the watershed algorithm and surveys the use of fuzzy relation to it. Section 3 introduces the proposed watershed algorithm using the composition of fuzzy relation. Section 4 shows the experimental results of the proposed algorithm. Section 5 concludes and discusses the experimental results of the proposed algorithm. Finally a list of references is given.

2. The Watershed Algorithm

The watershed algorithm [1,2,10,11] is very well suited for the problem of segmenting the different spots in a 2-D gel images, because after applying a small mean-filter, these spots are characterized by a monotonic increasing and thereafter decreasing shape. In this way it is possible to detect the catchment basins belonging to the different gel spots, see figure 2.

This is a very robust approach: varying background intensity has no influence on the finding of the different spot regions. To exclude small regions corresponding to background noise, a threshold was chosen for the minimal size of the basins. The remaining basins delineate the regions of most spots.

However, some spots overlap in such a way that they give rise to only one catchment basin, and as a result they will be identified as one spot.

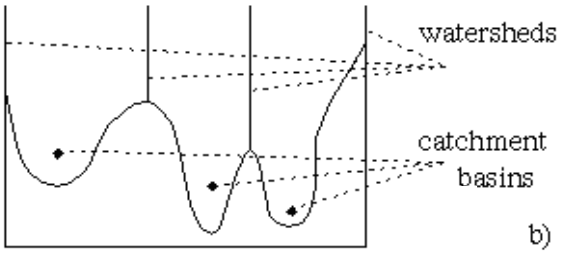

Figure 2.Watershed segmentation-Local minima yield catchment basins and local maxima define the watershed lines

To segment the spots from the background, the density peaks in the image have to be found.

A big advantage of this algorithm is that it is robust in the sense that it is not influenced by a variable background (low-frequency variations).

The watershed algorithm is a very robust for detecting spots, with the major advantage that there is no need for a background subtraction. Regarding this, the major disadvantage of the algorithm which is the over-segmentation must be overcome. In [3], Patino had suggested to the problem of watershed over-segmentation the use of the composition of three fuzzy relations $R_1 \circ R_2 \circ R_3$:pixel x_i is connected to pixel x_j and x_j has a gray value y and y belongs to cluster Z .He applied the proposed algorithm to standard images. He also used the Fuzzy c-means algorithm to "segment" the image before applying the watershed segmentation algorithm which had the disadvantage of increasing the complexity of the algorithm without using the benefits of the watershed algorithm. The concepts of fuzzy relations have been applied to the problem of watershed over-segmentation [3]. For this purpose let us define X as the set of m mosaic regions (or catchment basins) that we seek to simplify. Y is the set of n grey value levels from the histogram of the mosaic image. It is possible then to establish the two following relations:

R^1: x is connected to x,

R^2 : x has grey value y.

The first relation is defined by the following membership function:

$$\mu_{R_1}(x_i, x_j) = \begin{cases} 1 & if \quad i = j \\ 0 & if \ x_i \overset{\times}{\longleftrightarrow} x_j \\ \left| gv(x_i) - gv(x_j) \right| & if \ x_i \longleftrightarrow x_j \end{cases} \quad (1)$$

where gv(x) is a function giving the grey value level of region x,

and $x_i \longleftrightarrow x_j$ reads x_i is adjacent to x_j,

and $x_i \overset{\times}{\longleftrightarrow} x_j$ reads x_i is not adjacent to x_j

The matrix μ_{R_1} is actually a compatibility-relation matrix as its elements verify the properties of reflexivity $\mu(x_i, x_i) = 1$ and symmetry $\mu(x_i, x_j) = \mu(x_j, x_i)$.

The second relation is actually a crisp relation where

$$\mu_{R_2}(x_i, y_j) = \begin{cases} 1 & if \quad gv(x_i) = y_j \\ 0 & elsewhere \end{cases} \quad (2)$$

At this point, Patino[3] considered again the idea of the Fuzzy C-means algorithm employed to group together pixels having similar grey-level values [9]and he applied the composition of the fuzzy relations defined above to constrain merging the pixels only between adjacent regions.

Let Z be the set of c different clusters into which we want to simplify the mosaic region. We can define the following relation:

R_3: y belongs to z.

In this case the notion of belonging is that of a typical fuzzy set with $\mu_{R_3}(y_j, z_k)$, evaluating the membership of grey value y_j into cluster Z_k. The fuzzy matrix μ_{R_3} is found by running the Fuzzy C-means algorithm.

The choice of this particular clustering technique was made not only because it has already been implemented in a wide range of applications with successful results, but more particularly because it naturally leads to a fuzzy partition of the data whereas using other popular clustering methods such as the Leader algorithm (Hartigan, 1975) [4], Self Organizing Maps (Kohonen, 1998)[5], Substractif clustering (Yager and Filev,1994)[6], etc., would then need a further fuzzification step.

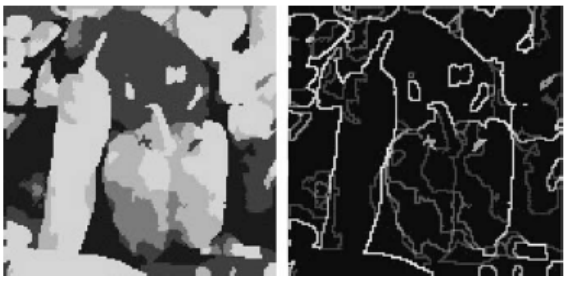

Fig. 3 Simplified mosaic and watershed of peppers images after connected regions are merged by the Patino's proposed approach

The whole procedure of watershed simplification has been reduced to the application of the following composition rule:R1 ∘ R2 ∘ R3: x_i is connected to x_j and x_j has grey value y and y belongs to cluster Z.

3. The proposed watershed algorithm using the composition of Fuzzy Relations

In our work, we will present a fully automatic 'watershed algorithm' for segmenting the image into different spot regions: the algorithm finds the appropriate regions so that in each region found, a spot can be quantified in a correct manner.

We will use only two relations. $R_1 \circ R_2$. The advantage of using only two relations is the ease of the algorithm and the elimination of redundancy.

Moreover, the connectivity is an important parameter for the watershed algorithm and adjusting this parameter before applying the algorithm has the intention of advancing the algorithm and reducing error, see figure 4. The two relations are defined as follows

R1: x_i has a 3×3 neighborhood x_j,

R2: x_j has grey value y such that y belongs to cluster Z.

x_j	x_j	x_j
x_j	x_i	x_j
x_j	x_j	x_j

Figure 4 shows the 3×3 neighborhood pixels

R1 is a fuzzy relation defined as follows

$$\mu_{R_1}(x_i, x_j) = \left| gv(x_i) - gv(x_j) \right| \quad (3)$$

Where x_j is a 3 ×3 neighbor of x_i.

Z is the set of clusters initially partitioned as follows $Z_1, Z_2, ..., Z_n$ each cluster contains 256/n point defined as Z^i is from $i*256/n$ to $(i+1)*256/n$ where i takes the values 0,1,…,n

R_2 is a crisp relation defined as follows

$$\mu_{R_2}(x_j, Z) = \begin{cases} 1 & if \quad gv(x_i) \in Z \\ 0 & elsewhere \end{cases} \quad (4)$$

The whole procedure of watershed simplification can be reduced to the application of the following composition rule:

R1 ∘ R2: x_i is connected to x_j and x_j belongs to cluster Z.

In our approach, there is no need to apply the Fuzzy C-means algorithm and the labeling is taken by maximizing the degree of membership values over all clusters i.e

$$x_{new} = \max_{Z_i = Z_1}^{Z_n} \mu_{R1 \circ R2}(x_{old}, z_i) \quad (5)$$

Since the second relation R_2 is a crisp relation then the max-min composition is equivalent to the max-product.

We can present the algorithm as in the following steps:

Step 1: Initialize clusters Z

Step 2: For each pixel find

$$\min_{x_j neighbor x_i} (\mu_{R_1}(x_i, x_j), \mu_{R_2}(x_j, Z))$$

where $\mu_{R_1}(x_i, x_j)$ and $\mu_{R_2}(x_j, Z)$ are computed by equations 3 and 4 respectively

Step 3: Labeling Pixels by applying equation 5

Step 4: Apply the watershed algorithm to the resulted mosaic images

4. Experimental Results

The LECB 2-D PAGE gel images database is available for public use. It contains data sets from four types of experiments with over 300 gif images with annotation and landmark data in html, tab-delimited and xml formats. It could be used for samples of several types of biological materials and for test data for 2D gel analysis software development and comparison with other similar samples.

PAGE is polyacrylamide gel electrophoresis. The LECB was the U.S. National Cancer Institute's Laboratory of Experimental and Computational Biology. Since this work was done, LECB has been reorganized as the CCR Nanobiology Program. The database is available at two Web sites [7,8]. In our work we used these data and applied our algorithm to 2D gel images of autologous human lymphoblastoid cell lines. The results are shown in the following figures.

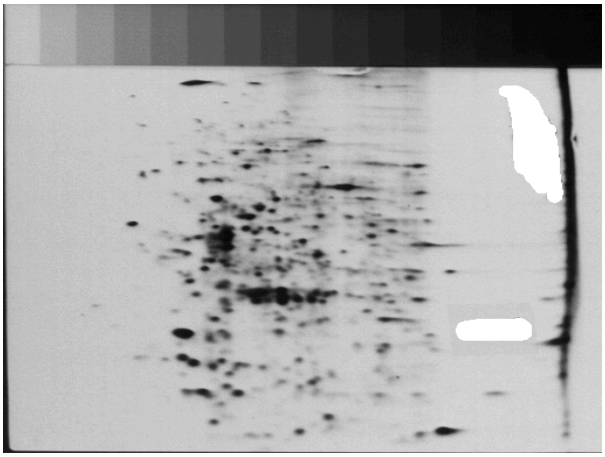

Fig 5. 2-D gel electrophoresis image of a Patient- Human leukemias

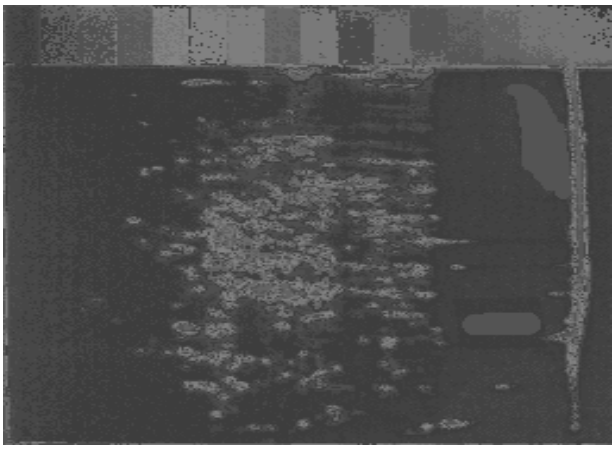

Fig 6. The Gradient image of 2-D gel electrophoresis image of a Patient-Human leukemias in fig. 3

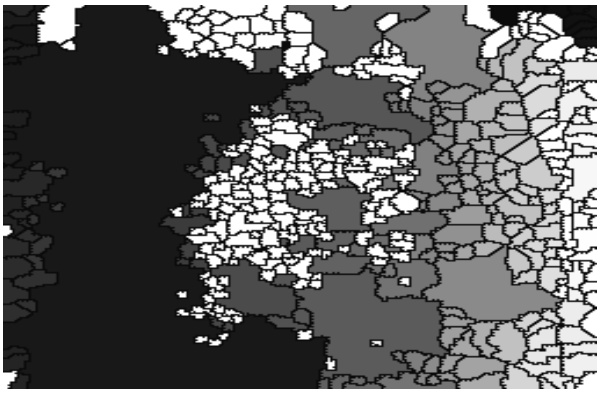

Fig 7. The gradient image of 2-D gel electrophoresis image after applying the Watershed algorithm

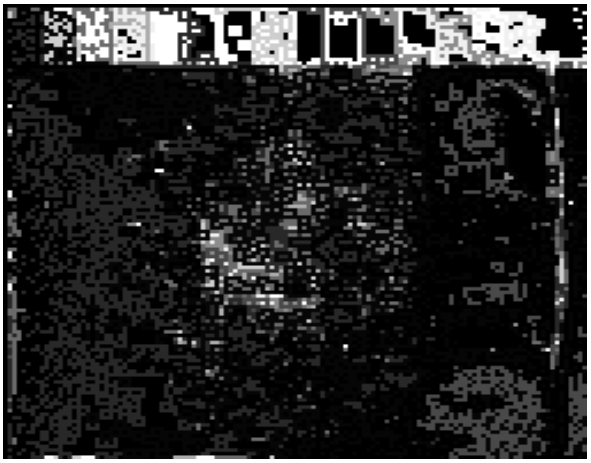

Fig 8. The mosaic simplified image of the gradient image of 2-D gel electrophoresis image after applying the composition of fuzzy relations

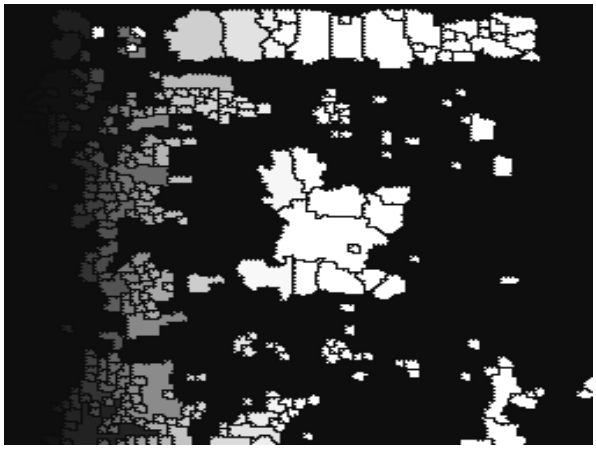

Fig 9. The Gradient image of 2-D gel electrophoresis image after applying the Fuzzy- Watershed Segmentation algorithm

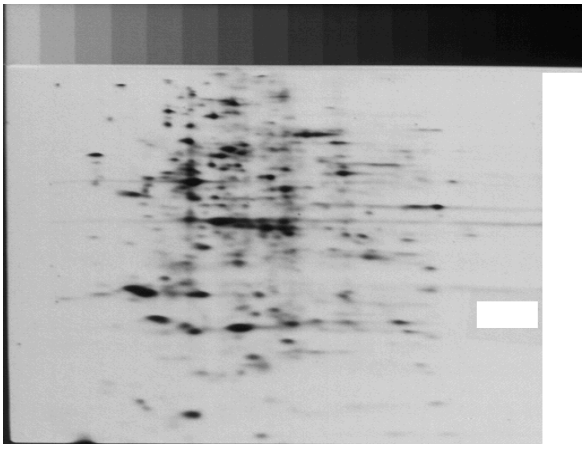

Fig 10. 2-D gel electrophoresis image of a second Patient- Human leukemias

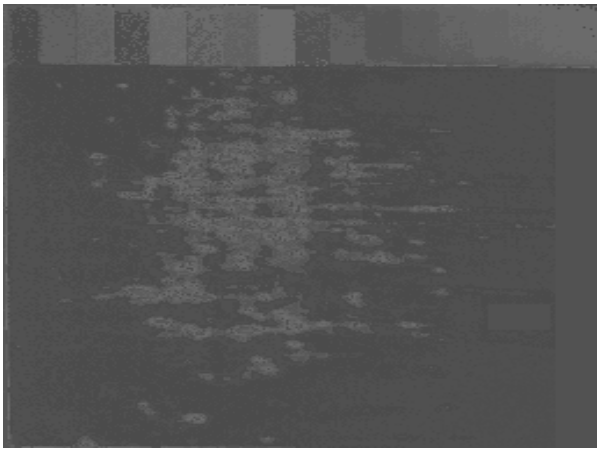

Fig 11. The gradient image of 2-D gel electrophoresis image of a second Patient- Human leukemias

Fig 12. The gradient image of 2-D gel electrophoresis image after applying the Watershed algorithm

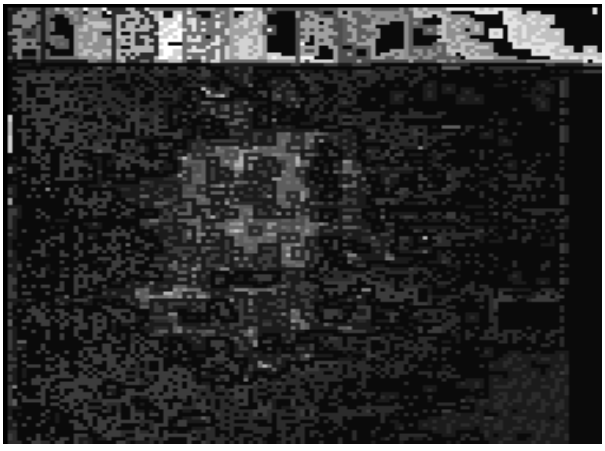

Fig 13 The mosaic simplified image of the gradient image of 2-D gel electrophoresis image after applying the composition of fuzzy relations

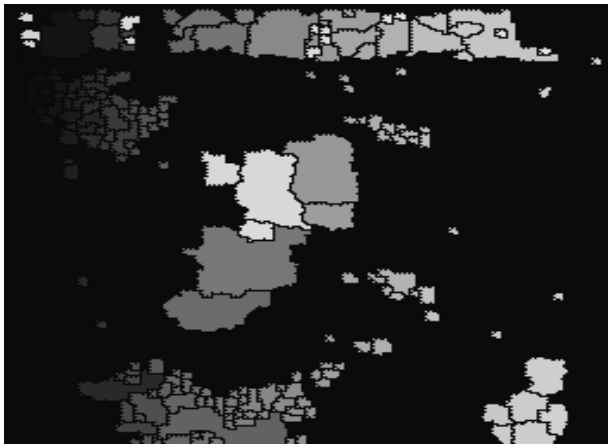

Fig 14. The Gradient image of 2-D gel electrophoresis image after applying the Fuzzy- Watershed Segmentation algorithm

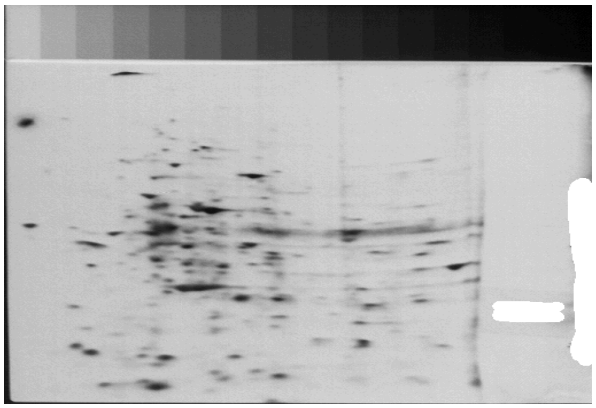

Fig 15. 2-D gel electrophoresis image of a third Patient- Human leukemias

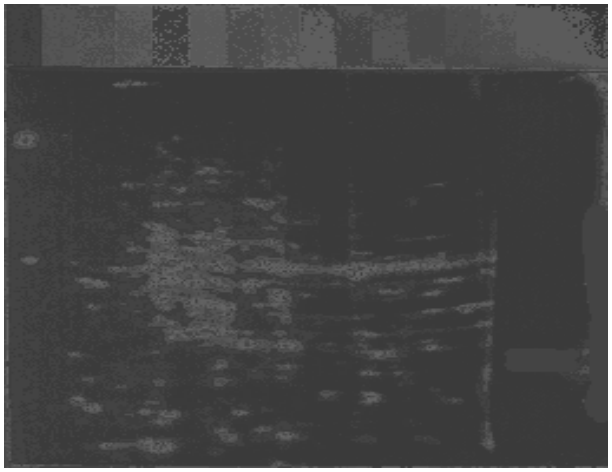

Fig 16. The gradient image of 2-D gel electrophoresis image of a third Patient- Human leukemias

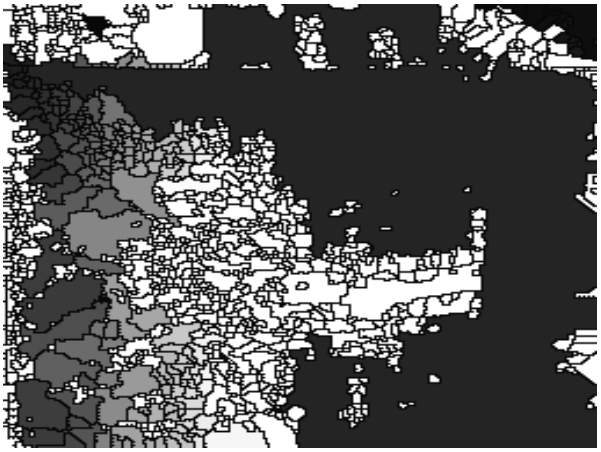

Fig 17. The gradient image of 2-D gel electrophoresis image after applying the Watershed algorithm

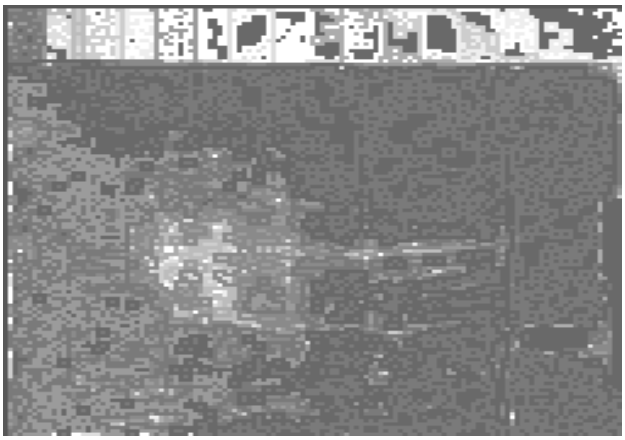

Fig 18. The mosaic simplified image of the gradient image of 2-D gel electrophoresis image after applying the composition of fuzzy relations

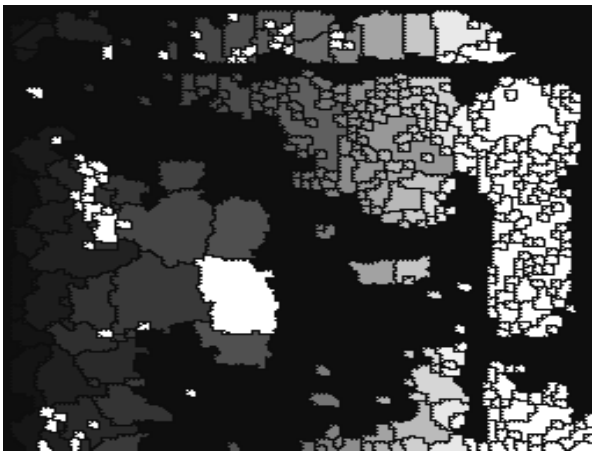

Fig 19. The Gradient image of 2-D gel electrophoresis image after applying the Fuzzy- Watershed Segmentation algorithm

5. Discussion

In this work, we presented a new algorithm based on the notion of fuzzy relations to segment and detect protein spots in 2-D gel electrophoresis images.
This algorithm shows high performance and detects the protein spots precisely. As illustrated in the figures from figure 5 to 19, the new algorithm simplifies the original image to a mosaic image where applying the watershed algorithm, the number of catchment basins is reduced and hence the problem of over-segmentation is handled. Comparing figures 7 and 9, 11 and 14, 17 and 19, we can say that the new algorithm succeeded in reducing the over-segmentation and identify the area where spots exist more precisely. The watershed algorithm has the advantage of partitioning the original images from the varying background which makes it suitable for our type of images. For future work, we suggest the development of fuzzy relations to obtain better results. The second relation can be a fuzzy relation defining the degree of membership of the grey value to a particular cluster for enhancement and improvement of the new algorithm.
Moreover, the use of intuitionistic fuzzy relations and its composition to partition the original image to mosaic regions. A key issue in this algorithm is the threshold of the minimal size of the basins. The estimation of this threshold for better results is a direction for future work.

Acknowledgment

This work had been supported by Mubarak City for Science and Technology

References

[1] E.Bettens, P. Scheunders, J. Sijbers, D.Van Dyck, L.Moens." Automatic Segmentation and Modeling of two-dimensional Electrophoresis gels ", in Proceedings of International Conference on Image Processing, 1996.
[2] Ming Hung Tsai et al., "Watershed-Based Protein Spot Detection in 2DGE Images", in the proceeding of ICS 2006, International Workshop on Software Engineering, Databases, and Knowledge Discovery, Taiwan, 2006
[3] Luis Patino, "Fuzzy relations applied to minimize over segmentation in watershed algorithms", Pattern Recognition Letters 26 (2005) 819–828
[4] Hartigan, J.A., 1975. "Clustering Algorithms". Wiley, New York.
[5] Kohonen, T., 1998. "The self-organizing map".Neurocomputing 21, 1–6.
[6] Yager, R.R., Filev, D.P., 1994. "Approximate clustering via the mountain method" . IEEE Transanctions on Systems Man Cybernetics. B 24, 1279–1284.
[7] www.ccrnb.ncifcrf.gov/2DgelDataSets
[8] bioinformatics.org/lecb2dgeldb
[9] Valente de Oliveira, W. Pedrycz "Advances in Fuzzy Clustering and its Applications" published 2007
[10] Beucher, S., Lantuéjoul, C., “Use of watersheds in contour detection”, Proc. Int. Workshop Image Processing, Real time edge and motion detection/estimation, Rennes, France, Sept. 17-21, 1979.
[11] Beucher, S., “Watersheds of functions and picture segmentation”, Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, Paris, France, May 1982, pp.1928-1931.
[12] Minh-Tuan Trong Hoang, Yonggwan Won, "A Marker-Free Watershed Approach for 2D-GE Protein Spot Segmentation," isitc, pp.161-165, 2007 International Symposium on Information Technology Convergence (ISITC 2007), 2007
[13] D.T. Lin and J.L. Kuo, "Improved Watershed Algorithm Spot Detection on Protein 2D Gel Electrophoresis Images", Proceeding (444) Signal and Image Processing - 2004
[14] E.Bettens, P.Scheunders, J.Sijbres, D.Van Dyck, and L.Moens,"Automatic Segmentation and modeling of two-dimensional Electrophoresis Gels," Proc. IEEE Int'l conf.on Image Processing, vol. 1, pp. 665-668, Sep . 16-19,1996.
[15] Ming-Hung Tsai, Hui-Huang Hsu, Chien-Chung Cheng, "Watershed-Based Protein Spot Detection in 2DGE images", in Proc. Int'l Computer Symposium (ICS 2006) , vol. III, p.p. 1334-1338, Taipei, Taiwan, Dec. 4-6, 2006.

Shaheera Rashwan Ph.D student in computer science department, Faculty of engineering ,Tanta University. Currently , Eng. Rashwan is a research assistant in Informatics Research Institute , Mubarak City for science and Technology