Watershed-Based Protein Spot Detection in 2DGE Images

Ming-Hung Tsai[a], Hui-Huang Hsu[b], Chien-Chung Cheng[c]
[a, b] *Department of Computer Science and Information Engineering*
Tamkang University, Taipei, Taiwan, R.O.C.
[c] *Department of Applied Chemistry*
National Chia-Yi University, Chia-Yi, Taiwan, R.O.C.
E-mail: 693191560@s93.tku.edu.tw, h_hsu@mail.tku.edu.tw, cccheng@mail.ncyu.edu.tw

ABSTRACT

2D gel electrophoresis (2DGE) plays an important role in proteomics. It can separate proteins effectively with their pI values and molecular weights. Proteomics researchers needed to identify interested protein spots by examining the gel. This is time-consuming and labor extensive. It is desired that the computer can analyze the proteins automatically by first detecting and quantifying the protein spots in the digitized 2DGE images. In this paper, we re-investigate the use of the watershed algorithm in segmenting the protein spots from the varying background. However, the watershed algorithm often produces an over-segmented result. Thus, marker-based methods are used to improve the segmentation result. The result is compared with that generated by a commercial software tool – ImageMaster. It is demonstrated that our system can detect protein spots more precisely.

1: INTRODUCTION

The two-dimensional gel electrophoresis (2DGE) technique plays an important role in proteomics, because it can effectively separate the proteins in a cell according to their iso-electric point (pI) values and molecular weights (Mr). The result of 2DGE is many dark spots on the gel [1]. Each spot represents a protein or a group of proteins. The proteomics researcher has to check the gel carefully in order to identify interested proteins for further analysis. There are usually hundreds or even thousands of protein spots on a gel. So this could slow down the whole analysis process. And computerized analysis on the gel image is desired.

To analyze protein spots on the 2DGE image, segmentation of the spots from the varying background is the first task. Then quantitative analysis of each segmented spot becomes possible. Each segment is represented by its boundary. Finally, image registration techniques are used in comparing two related 2DGE images to identified interested protein spots. In this paper, we concentrate on the first two parts and watershed algorithms are utilized.

In [2], 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. Little discussion is made for the watershed-based segmentation and the segmentation result is not that good. In [3], a new approach was proposed for finding similar proteins in 2DGE images. The paper mainly addresses issues of 2DGE image registration mentioned above.

For segmentation of protein spots from the 2DGE image, the first step is to smooth the image. The reason is that the 2DGE image is often noisy. The next step is to segment protein spots from the background. Because the background of 2DGE images is uneven, setting a simple threshold is not sufficient to segment the spots. More sophisticated image segmentation techniques are needed. They can be roughly divided into the following three categories [4].

(1) Edge-based methods: The goal of the methods in this category is to detect the object edges, and then group the edges into object contours. Filters can be used here. For example, the Laplacian filter.
(2) Region-based methods: These methods can divide the image into several small regions according to the predefined criteria. The iterative merging process is then used to merge similar neighboring regions. Until the status becomes stable, the merging process stops.
(3) Hybrid methods: Methods of this category provide an optimal solution for image segmentation by combining the previous two methods. For example, after image segmentation by the watershed algorithm, the over-segmentation problem occurs. So region merging is necessary for avoiding too many small regions and producing meaningful edges for objects.

After image segmentation for detecting protein spots,

the next step is quantification of the protein spots. Data like the coordinates, the area, and the volume of each spot can be collected easily from the segmentation result. These data are important for proteomics research. Our system also calculates and saves the data in a database table for each 2DGE image. And the detected spots are numbered so that the subsequent registration is made easier.

The rest of the paper is organized as follows. Section 2 introduces the watershed-based methods used in our system, including the traditional method and the marker-based method. Section 3 presents the experimental results. In Section 4, a brief conclusion is drawn and the future work is discussed.

2: WATERSHED ALGORITHMS

The watershed transform is an important method in morphological image processing. It is good in precisely segmenting objects from the background. The idea is that the image is considered as a map and the gray level of the image is a topographic relief. The algorithm can build watersheds between neighboring catchment basins and these watersheds are generally the boundaries of objects [4][5][6]. Watersheds can be built by flooding simulations. It is assumed that a hole is punched at each local minimum. The whole 3D map is then immerged into water. The water will slowly flood all catchment basins. Also, dams are built when waters from two adjacent catchment basins would merge. After the flooding is finished, the dams built represent the watersheds. And the watersheds give a partition of the source image.

The detection of protein spots using the watershed algorithm can be divided into the following four steps:
(1) Noise reduction for the source gel image,
(2) Generation of the gradient image,
(3) Application of a threshold on the gradient image,
(4) Watershed segmentation .

Since the gel image is often noisy, in the first step, existing filters are used to remove the noise, for example, the median filter and the low-pass filter. This step can make the gel image smoother. The second step is to calculate the gradient value of the gel image. In this step, the Sobel filter or morphological gradient can be used [7]. A gradient operator is applied to calculate two directional derivatives of the image and the magnitudes are subsequently combined. Figure 1 shows the source 2DGE image. Figures 2 and 3 are the resulted images of Step1 and Step 2, respectively. It can be seen that the gradient image gives a general idea about the locations of protein spots. But the boundaries are not precise enough.

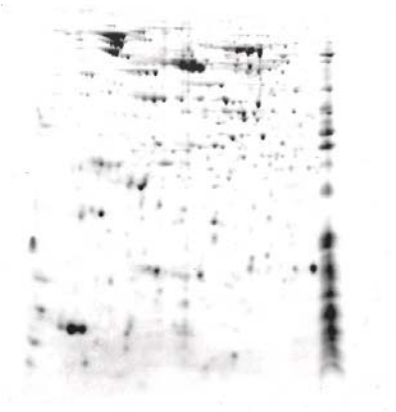

Figure 1. Source 2DGE image

Figure 2. Result after noise reduction

Figure 3. Gradient image of 2DGE image

In Step 3, a threshold (T) value is used to filter the gradient value. If the gradient value of a pixel is less than the threshold, it is modified to zero. This is because the gradient value represents the variation between pixels. When variation is less than the threshold, it can be

regarded as noise and then filtered. The procedure is defined by

$$G(x)=\begin{cases}0, & \text{if } G(x)<T\\ G(x), & \text{else}\end{cases}$$

Finally, watershed transform [5] can be used to segment the protein spots from the background. After the gradient image is applied with a threshold, the traditional watershed method can detect protein spots from the image easily. However, the result is usually not satisfiable. Because taking threshold of the gradient image can not filter out all noise, the result of this method could be left with incorrect protein contours. Also, over segmentation is still a problem. The result is shown in Figure 4. Although the result seems alright, many undesired spots are detected.

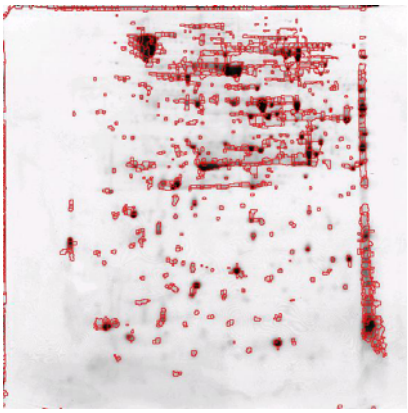

Figure 4. Segmentation Result by traditional watershed method

Our system uses the marker-based watershed algorithm to produce a more accurate protein contour. It is well-known that over-segmentation is a big problem of the watershed method due to noise. The image has so many local minima. And it is very possible that the protein spots locate on them. So it is desired that the minima on the proteins are preserved and those on the noise are removed. Minima imposition can be used to reach this goal. It is a filtering technique and is used to remove spurious minima. Using minima imposition needs a marker function that marks the relevant objects (inner markers) and their backgrounds (outer markers). Thus the marker-based method can be divided into four steps [8][9].

(1) Inner markers determination,
(2) Outer markers determination,
(3) Minima imposition,
(4) Watershed segmentation.

First, we must find the inner markers. Because the protein spots which we want to find out are dark against to the background, we can use this feature to mark them. The inner marker is very important because it represents the objects which we want to segment. So we hope to mark them as many as possible. Morphological reconstruction [8] is a good method which can help us to carry it out because the intensity feature of proteins reveals that they are local minima. We can add height *h* to pixels on the source gel image and use this image as the marker image. And then reconstruction by erosion of the marker image from the mask image is performed.

The next step is to find the outer markers. An outer marker defines the scope of the corresponding inner marker. So we can use the watershed algorithm of inner markers to find their corresponding outer markers. Inner markers and outer markers are then combined to produce the marker image. The marker image is defined by

$$f(x)=\begin{cases}0, & \text{if } x \text{ belongs to marker}\\ \text{Tmax}, & \text{otherwise}\end{cases}$$

The third step is to remove unsignificant minima by minima imposition. The gradient image instead of the source image is used here. In order to get more precise contours of the proteins, modification of the gradient image is necessary. Point-wise minimum of the gradient image and the marker image produces the needed mask image. And point-wise minimum of the gradient image plus one and the marker image produces the needed marker image. Reconstruction by erosion is then performed using the obtained mask image and marker image. The minima on the marker image can be reserved and others can be filled. So we can reserve the desired minima and remove the irrelevant. In this way, the over-segmentation problem can be solved because the unnecessary minima are removed. Finally, watershed transform can be used to segment the protein spot contours and the over-segmentation problem is no longer a problem.

3: EXPERIMENTAL RESULTS

In this section, the preliminary results for detecting the protein spots in a 2DGE image by the marker-based watershed method mentioned in Section 2 are presented. The system is implemented by Borland C++ Builder 6. First, in our system, the main function is detection and segmentation of protein spots. Secondly, quantification data of the proteins, for example, area and volume, can be provided.

The first step is to find the inner markers and the outer markers. Figure 5 shows the resulted inner markers. Figure 6 shows the resulted outer markers in the gel image. The outer marker is produced by applying watershed transform on the inner marker image. Then the marker image and the gradient image are used for minima imposition. In Figure 7, we can see that the minima on protein spots are preserved and others are removed. So the

over-segmentation problem on the subsequent watershed transform can be avoided. Finally, we can get the final result in Figure 8. It is produced by applying the watershed transform on Figure 7. Every protein spot in the gel image can be detected and its contour is very clear and precise. In this gel image, 273 protein spots are detected. All the protein spots are numbered and saved along with the quantitative data. Finally, quantitative data of each protein spot can be displayed in a small window, as shown in Figure 9.

The result by our system is compared with the result by a commercial software tool – ImageMaster [10]. Figure 10 shows the segmentation result of the same gel image by ImageMaster. Totally, 569 protein spots are detected. The number is more than double comparing with the result by our system in Figure 8. Although our system also wrongly detects a few spots, it is not as serious. Furthermore, bizarre protein contours can be seen at the lower part of Figure 10. Figure 11 shows the details of this part. On the contrary, the spot locations of our result are right on target. And, the contours of proteins of our result are more precise than those by ImageMaster.

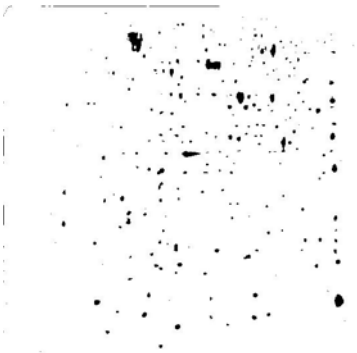

Figure 5. Inner markers

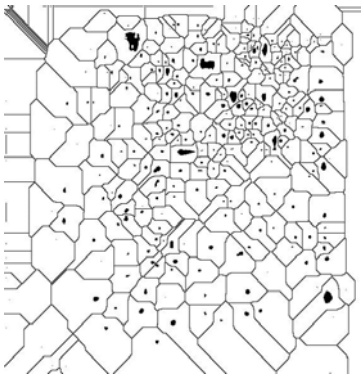

Figure 6. Outer markers

Figure 7. Minima imposition

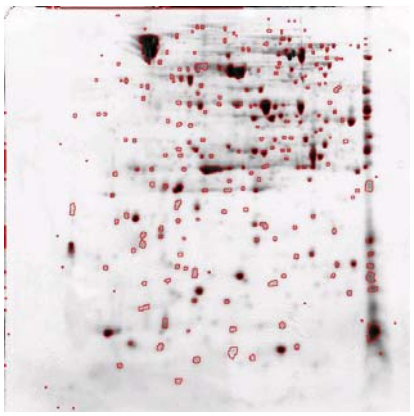

Figure 8. Segmentation result

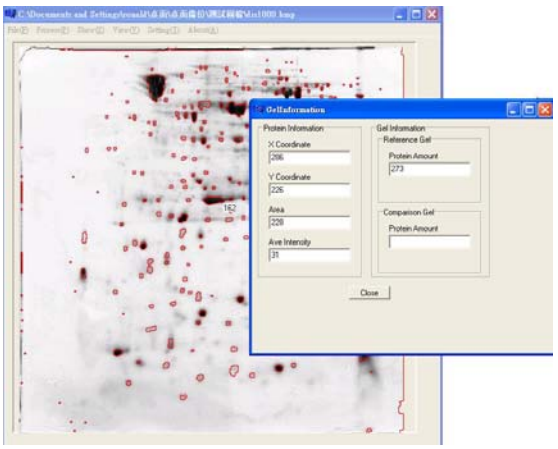

Figure 9. Quantitative data of a protein spot

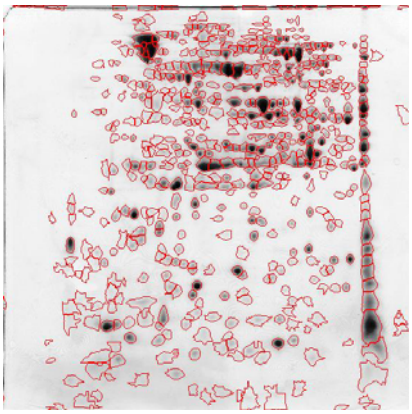

Figure 10. Segmentation result by ImageMaster

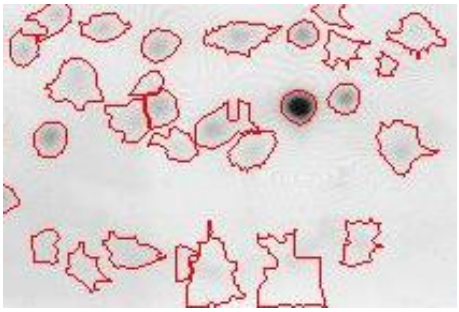

Figure 11. Selected unusual protein spots in the result by ImageMaster

4: CONCLUSIONS AND FUTURE WORK

We have developed a system which can detect and quantify protein spots in 2D gel electrophoresis images. The system results in a more precise segmentation than ImageMaster. The segmentation lines fall exactly on the boundaries of the protein spots. However, there are still problems to be conquered. For example, overlapping proteins cannot be segmented well. When our system processes these spots, the result is often just one spot. But in fact, there are more than one proteins. The bio-technician would cut such spots from the gel and reprocess the multiple proteins on one spot with a higher resolution. So in a new 2DGE experiment, the proteins can be separated on the gel. Also, the result of our system is sensitive to the threshold setting. It is sometimes hard to tune the thresholds. For the future work, automatic matching of similar protein spots on two different 2DGE images is also desired. With this registration function, 2DGE images from the same experiment can be easily compared. And changes of protein distribution can be detected.

REFERENCES

[1] D. E. Krane and M. L. Raymer, *Fundamental Concepts of Bioinformatics*, Benjamin Cummings, San Francisco, CA, USA, 2003.

[2] E.Bettens, P. Scheunders, J. Sijbers, D. Van Dyck, and L. Moens, "Automatic Segmentation and Modellng of Two-dimensional Electrophoresis Gels," *Proc. IEEE Int'l Conf. On Image Processing*, vol. 1, pp. 665 – 668, Sep. 16 – 19, 1996.

[3] Nawaz Khan and Shahedur Rahman, "A New Approach To Detect Similar Proteins from 2D Gel Electrophoresis Images," *Proc. of the Third IEEE Symp. on Bioinformatics and Bioengineering* (BIBE'03), pp. 182 – 189, March 10 – 12, 2003.

[4] K. Haris, S. N. Efstratiadis, N. Maglaveras, A.K. Katsaggelos, "Hybrid Image Segmentation Using Watersheds and Fast Region Merging," *IEEE Trans. on Image Processing*, vol. 7, no. 12, pp. 1684 – 1699, Dec. 1998.

[5] Luc Vincent and Pierre Soille, "Watersheds in Digital Spaces: An Efficient Algorithm based on Immersion Simulations," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 13, no. 6, pp. 583 – 598, June 1991.

[6] D. Hagyard, M. Razaz, and P. Atkin, "Analysis of Watershed Algorithms for Greyscale Images," *Proc. of Int'l Conf. on Image Processing*, vol.3, pp. 41-44, Sep. 16 – 19, 1996.

[7] J. C. Russ, *The Image Processing Handbook*, CRC Press, 1999.

[8] Luc Vincent, "Morphological Gray scale Reconstruction in Image Analysis: Applications and Efficient Algorithms," *IEEE Trans. on Image Processing*, vol. 2, no. 2, April 1993.

[9] Pierre Soille, *Morphological Image Analysis: Principles and Applications*, Berlin: Springer, 1999.

[10] ImageMaster 2D Platinum v6.0, http://www.amershambiosciences.com, location: Home > Products > Proteomics > Image Analysis > ImageMaster 2D Platinum v6.0, last accessed on 2006/8/8.