21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

2D-GE Image Analysis Focusing on Elimination of Spurious Spots

Eirini Kostopoulou[a], Eleni Zacharia[a], Dimitris Maroulis[a], and Antonia Vlahou[b]
[a]*Department of Informatics and Telecommunications, University of Athens, Greece*
[b]*Biomedical Research Foundation of the Academy of Athens, Greece*
ikostop@di.uoa.gr, eezacharia@gmail.com, dmaroulis@di.uoa.gr, vlahoua@bioacademy.gr

ABSTRACT

2D Gel Electrophoresis image analysis is widely recognized as one of the most crucial processes following a proteomic experiment. Amongst its stages, detection and segmentation are the most challenging ones. The available software packages and techniques fail to detect and segment some of the real spots while they often detect a vast number of spurious spots. In this paper, an original approach to analyze gel electrophoresis images is presented. The proposed approach is an improvement on our earlier version because it tackles the aforementioned deficiencies of existing software programs and techniques in a more efficient manner. The experiments conducted on a set of 16-bit 2D gel electrophoresis images, demonstrate that it outperforms state-of-the-art methods. Indeed, it detects more real protein spot, as well as eliminates spurious spots.

1. Introduction

2D Gel Electrophoresis (2D-GE) is a well-established proteomics technique used for the separation of complex mixtures of proteins [1]. It has been utilized in a wide variety of different application areas such as cancer research, infectious disease diagnosis and treatments as well as pharmacology research [2].

The output of the 2D-GE technique is a digital grayscale image containing thousands of protein spots. The process of analyzing a 2D-GE image can be divided into three main stages [3]: spot-detection, spot-segmentation and spot-quantification. Amongst them, the detection and segmentation stages are the most crucial and challenging ones as they considerably affect the final biological conclusions of the proteomic experiment. The main reason behind this lies in the nature of 2D-GE images. Indeed, these images contain thousands of spots of various intensities, sizes and shapes. In many cases, spots are so poorly contrasted that they are not clearly visible. Furthermore, adjacent spots often overlap and are therefore hard to be separated. Finally, the quality of these images is often degraded due to the existence of noise, artifacts as well as inhomogeneous background [3].

Several software packages for 2D-GE image analysis are available and used in biologic laboratories, (i.e. Delta 2D [4], Melanie 7 [5]). Despite their respective merits, they cannot detect all the real spots and they detect an enormous number of spurious spots.

According to [6] accurate results stemming from the existing software packages and techniques have not been achieved up to date. Watersheds [7] have been introduced to cope with the presence of overlapping spots. However, they lead to over-segmentation. Inner-marked watershed methods [8] cope with over-segmentation; nevertheless, they require user intervention. The active contour-based method [9] tackles most challenges; however, it cannot segment overlapping spots. Anjos et al. method [10] -called "*Scimo*"- is based on watersheds. Although, it confronts most of the challenges, it fails to separate highly overlapping spots and it produces rough spots boundaries. As a result, human intervention is required for selection of mandatory input parameters as well as correction of the results. However, human intervention is a time-consuming procedure and adds subjectivity to the results.

Our previous work [11] includes two steps: Firstly, regions of interest (ROIs) that contain mostly spots are detected. Subsequently, these ROIs are smoothed with a median filter and segmented based on their local intensity maxima. However, it resulted in: a) over-segmented spots, due to the detection of multiple local intensity maxima within a single spot (Fig. 1a), b) spurious spots, due to the detection of local intensity maxima in regions that contain background or artifacts (Fig. 1b, c).

In this paper, an original approach for spot detection and segmentation of 2D-GE images is presented. The proposed approach, builds upon our earlier approach by improving the process of determining the spot-centers as well as by deleting spurious spots. Several experiments were conducted on real 2D-GE images containing ~1000 spots in order to evaluate the proposed approach. The obtained results demonstrate that the proposed approach is effective and it outperforms state-of-the-art commercial software packages and techniques.

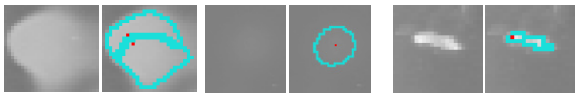

Figure 1. (a) Over-segmented spot. Spurious spots containing: (b) background, (c) artifact.

2. Proposed Approach

2.a. Detection of ROIs that contain mostly spots

Initially, the 2D-GE image is roughly segmented into: 1) a set of regions *R* containing mostly spots (ROIs) and 2) the background region, as described in our earlier work [11]. Fig. 2b depicts in black the background and in gray scale the four regions of the *R* set (r_1, r_2, r_3, r_4) of Fig. 2a. According to the ground truth on this image, there are eleven spots, the location of which is illustrated with blue crosses (Fig. 2a). Seven spots are located in r_1, two in r_2 and one in each of r_3 and r_4. The next steps of the proposed approach are the segmentation of each ROI into spot regions based on groups of local intensity maxima and the deletion of the spurious spots as described below.

2.b. Segmentation of ROIs based on groups of local intensity maxima

The spot-centers located in each distinct region of the *R*-set correspond to local maxima of the intensity values of pixels. However, not all local maxima correspond to spot-centers as 2D-GE images are often contaminated with noise, artifacts and contain inhomogeneous background. Therefore, instead of determining the spot centers based on the local maxima in the median-intensity values of a window *w* of pixels, we define '*groups of local maxima*' (*GoLMs*) based on their intensity values and positions. The determination of the GoLMs is conducted as follows: For each distinct region *r* of the *R*-set, a graph $G_r=(V, E)$ is defined. The vertexes v_m, $m=1,\ldots,|V|$ of

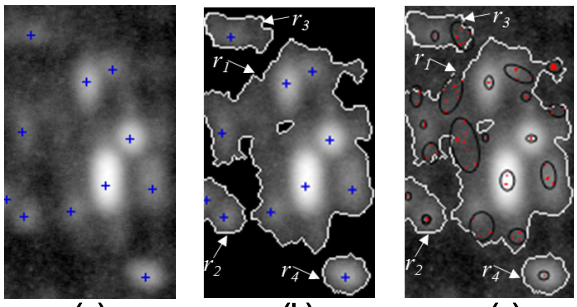

Figure 2. (a) A 2D-GE sub-image, (b) the R set of regions, (c) the local maxima arranged into the GoLMs

the G_r graph represent the pixels *p* of the following attributes: 1) *p* are located in the *r* region and 2) their intensity-values *I(p)* correspond to local maxima of the intensity values of the pixels. Two vertexes of the graph are connected only if the distance between the pixels they represent is less than T_d. As a result, the G_r graph contains a number of subgraphs, each considered to be a distinct *GoLM* that corresponds to a unique spot. For example, Fig. 2c depicts with red color the local maxima which are located on the *R*-set of regions of the 2D-GE sub-image of Fig. 2a. The black ellipsoids depict the *GoLMs*. The r_1 region contains 16 *GoLMs*, the r_2, r_3 two, and the r_4 only one. Subsequently, each region *r* of the *R*-set is further segmented into sub-regions around the determined *GoLMs*. Thus, a number of sub-regions equal to the number of distinct *GoLMs* is extracted. These segmented sub-regions contain either spot-pixels and background-pixels or only background-pixels. The optimal thresholding technique [12] is applied to each sub-region in order to extract the spot-surfaces *S*.

2.c. Erosion of spurious spots based on statistical measures

Fig.3a depicts the spots' surfaces stemmed from the previous step. On this image one may observe that a number of spot-surfaces *S* does not contain protein spots but spurious spots; in reality they are background regions contaminated with noise and artifacts. Based on empirical observations these regions contain either many local maxima or are extremely homogenous. As a result, spurious spots (i.e. spot-surfaces with yellow contour in Fig. 3a) are those which fulfill one of the following two conditions:

Condition 1: The percentage of local maxima inside a spot-surface is higher than a threshold T_m.

Condition 2: The coefficient of the variation CV [13] in the intensities values inside a spot-surface is lower than a minimum acceptable value T_{cv}. Indeed, if

the CV is lower than T_{cv} it means that the dispersion of the intensities' distribution is small, thus the spot-surface is in reality a background region. Fig. 3b depicts the final result of our approach, after the deletion of spurious spots. It becomes obvious that the proposed approach has segmented all the eleven real protein spots while avoiding the detection of any spurious spot.

3. Results

Several experiments were performed in order to evaluate the performance of the proposed approach and compare it with other outstanding techniques. The dataset used in these experiments is the Anjos et al. dataset which has been already been used for the evaluation of state-of-the art methods (i.e., Scimo).

It is worth pointing out that although we have evaluated our method on Anjos et al. dataset, the parameters T_d, T_m and T_{cv} of our method were experimentally adjusted based on the dataset utilized in [11] and their values were set to 8, 0.025 and 0.001, respectively. This proves that the parameters set works well on various sets of images. On the contrary, the parameters of the other commercial programs and techniques were adjusted – by expert biologists – according to each separate image. Therefore, the results of the proposed method were compared with the optimal results of the other commercial programs and techniques.

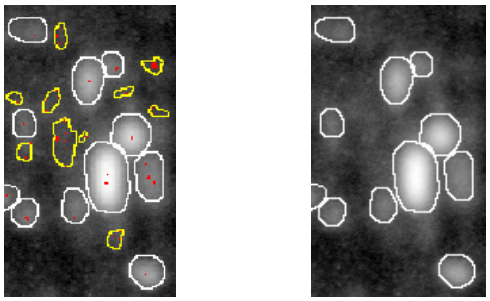

Figure 3. The 2D-GE sub-image of Fig. 2a containing the spots' surfaces before (a) and after (b) the erosion of spurious spots

The detection results were statistically evaluated using sensitivity and precision. This statistical analysis was based on the ground truth provided by expert

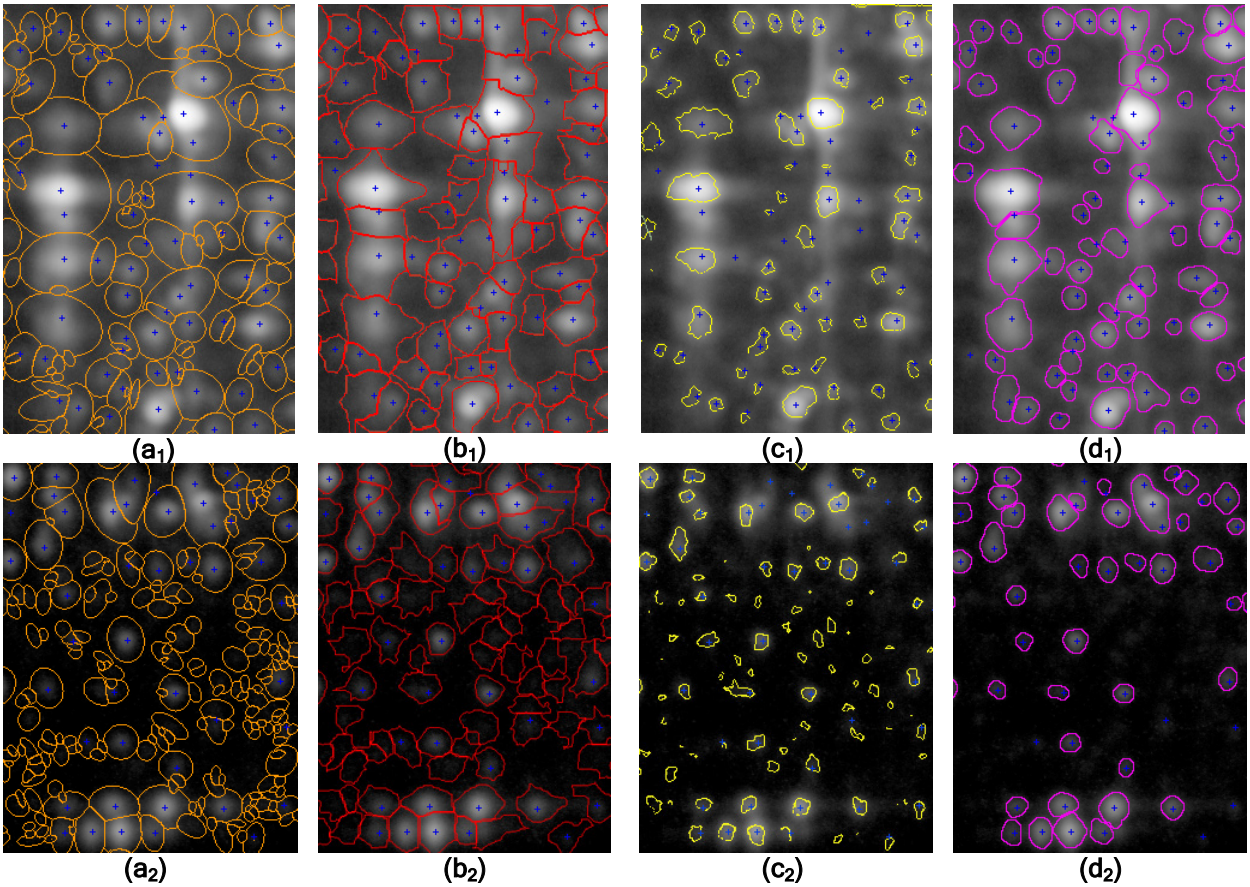

Figure 4. Detection and segmentation results produced by (a_1, a_2) Delta 2D, (b_1,b_2) Melanie 7, (c_1,c_2) Scimo and (d_1,d_2) the proposed approach.

Table 1. Comparison of detection results

Methods	Precision	Sensitivity
Melanie	81.7 ± 14.1	94.8 ± 1.1
Delta 2D	48.4 ± 17.8	88.2 ± 4.2
Scimo	75.0 ± 18.5	86.5 ± 2.4
Proposed approach	**97.2** ± 1.8	**94.3** ± 3.7

biologists of the Biomedical Research Foundation of the Academy of Athens who have manually determined the locations of the protein-spots by drawing a cross inside each unique spot-region. The segmentation results were optically evaluated by the same expert biologists, since segmentation ground truth was not available. Table 1 presents the comparison of detection results between the proposed approach, Delta 2D and Melanie 7 software packages as well as Scimo technique. It is worth mentioning that the higher the method's sensitivity value is, the fewer real spots are overlooked. Likewise, the higher the method's precision value is, the fewer spurious spots are detected. Table 1 makes clear that the proposed approach i) detects the protein spots almost perfectly, ii) detects the smallest amount of spurious spots, and iii) is radically more successful than the other three software programs and techniques.

The detection and segmentation results of the application of Delta 2D, Melanie 7, Scimo as well as of the proposed approach on two 2D-GE sub-images are illustrated in Fig. 4. Figs. 4(a_1-d_1) and Figs. 4(a_2-d_2) contain 74 and 48 spots, respectively. The location of each spot - according to the ground truth- is illustrated with a blue cross. On these images, one may observe that Delta 2D and Melanie 7 merge overlapping spots, segment spurious spots, and include background in the boundaries of real spots. Melanie 7 also misses faint spots. Scimo misses numerous faint spots, segments spurious spots, and fails to generate plausible boundaries. On the contrary, the proposed approach has effectively detected and segmented almost all of the protein spots.

4. Conclusions

In this paper, an original method for spot-detection and spot-segmentation of 2D-GE images is presented. It determines more accurately the spot-centers, and it avoids the detection of spurious spots. The experimental results over real 2D-GE images demonstrate that it is very efficient as it achieves 97.2% precision and 94.3% sensitivity. Furthermore, it outperforms software programs such as Melanie 7 and Delta 2D and state-of-the-art methods, such as Scimo.

5. Acknowledgements

The authors would like to thank the authors of Scimo for providing the dataset utilized in their experiments. This research has been financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: THALIS, UOA, CERVI-CAN-PROT.

6. References

[1] J.L. López, "Two-dimensional electrophoresis in proteome expression analysis," *Journal of Chromatography B*, vol. 849, no. 1-2, 2007, pp. 190-202.
[2] J.M. Lee, J.J. Han, G. Altwerger, and E.C. Kohn, "Proteomics and biomarkers in clinical trials for drug development", *Proteomics J.*, vol. 74, no. 12, 2011, pp.2632-2641.
[3] A.W. Dowsey, M.J. Dunn, and G.Z. Yang, "The role of bioinformatics in two-dimensional gel electrophoresis," *Proteomics J.*, vol. 3, no. 8, 2003, pp. 1567–1596.
[4] Delta2D [www.decodon.com]
[5] Melanie 7 [www.genebio.com]
[6] B.N. Clark, and H.B. Gutsein, "The myth of automated, high-throughput two-dimensional gel analysis," *Proteomics*, vol. 8, no.6, 2008, pp. 1197-1203.
[7] K.P. Pleissner, F. Hoffman, K. Kriegel, C. Wenk, S. Wegner, A. Sahlstrom, H. Oswald, H. Alt and E. Fleck, "New algorithmic approaches to protein spot detection and pattern matching in two-dimensional electrophoresis databases," *Electrophoresis*, vol. 20, 1999, pp. 755-765.
[8] Y. Kim, J. Kim, Y. Won, and Y. In, "Segmentation of protein spots in 2-D gel electrophoresis images with watershed using hierarchical threshold", *Computer and Information Sciences*, vol. 2869, 2003, pp. 389-396.
[9] M. Savelonas, E. Mylona, and D. Maroulis, "Unsupervised 2D Gel Electrophoresis Image Segmentation based on Active Contours," *Pattern Recognition*, vol. 45, no. 12, 2012, pp. 720-731.
[10] A dos Anjos, AL Møller, BK Ersbøll, C. Finnie, and H.R. Shahbazkia, "New approach for Segmentation and Quantification of Two-Dimensional Gel Electrophoresis Images," Bioinformatics, vol. 27, no.3, 2010, pp. 368-375.
[11] E. Zacharia, E. Kostopoulou E, D. Maroulis, and S. Kossida, "A Spot Segmentation Approach for 2D Gel Electrophoresis Images Based on 2D Histograms", Proc. 20th International Conf. on Pattern Recognition, pp. 2540-2543, Istanbul, Turkey, 2010.
[12] R. Gonzalez, R. Woods: *Digital Image Processing*. Addison-Wesley Longman Publishing Co., Boston; 2002.
[13] A. Hayter, *Probability and Statistics for Engineers and Scientists,* Duxbury Press, USA, 2011.