An Effective Approach for Detection and Segmentation of Protein Spots on 2-D Gel Images

Eirini Kostopoulou, Eleni Zacharia, and Dimitris Maroulis, *Member, IEEE*

***Abstract*—Two-dimensional gel image analysis is widely recognized as a particularly challenging and arduous process in proteomics field. The detection and segmentation of protein spots are two significant stages of this process as they can considerably affect the final biological conclusions of a proteomic experiment. The available techniques and commercial software packages deal with the existing challenges of 2-D gel images in a different degree of success. Furthermore, they require extensive human intervention which not only limits the throughput but unavoidably questions the objectivity and reproducibility of results. This paper introduces a novel approach for the detection and segmentation of protein spots on 2-D gel images. The proposed approach is based on 2-D image histograms as well as on 3-D spots morphology. It is automatic and capable to deal with the most common deficiencies of existing software programs and techniques in an effective manner. Experimental evaluation includes tests on several real and synthetic 2-D gel images produced by different technology setups, containing a total of ∼21 400 spots. Furthermore, the proposed approach has been compared with two commercial software packages as well as with two state-of-the-art techniques. Results have demonstrated the effectiveness of the proposed approach and its superiority against compared software packages and techniques.**

***Index Terms*—Two-dimensional (2-D) gel images, protein spot detection, protein spot segmentation, proteomics.**

I. Introduction

SCIENTIFIC interest in the field of proteomics has spectacularly increased in recent years. In this field of research, a number of opportunities has emerged; to investigate a multitude of diseases and to reach and extract useful conclusions for their treatment, to produce new drugs, to demonstrate new diagnostic markers and to explore biological events [1]–[4]. The 2-D Polyacrylamide Gel Electrophoresis technique has been widely used in proteomics because of its ability to separate thousands of proteins on polyacrylamide gels, according to the differences in their net charge and their molecular mass [5]–[7]. Digitized 2-D gels images contain thousands of spots, each representing a specific protein. Image analysis is therefore crucial in extracting biological information from a 2-D gel electrophoresis experiment. One of the goals of such an analysis is the rapid identification of 1) proteins located on a single gel and 2) differentially expressed proteins between samples run on a series of 2-D gels.

The process of analyzing 2-D gel images includes the following four main stages: 1) spot detection, 2) spot segmentation, 3) spot quantification as well as 4) image alignment for matching the corresponding protein spots in different images. The process of analyzing 2-D gel images can be performed in two different ways [8], [9]: in the customary analysis workflow, spot detection and segmentation are performed prior to image-alignment [10], [11] while in another analysis workflow, image-alignment is applied prior to spot detection and segmentation (Delta2D and Progenesis Samespots). Regardless of the workflow performed, spot detection and segmentation are two challenging stages in 2-D gel image analysis due to the very nature of these 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 are often not separated; instead they are highly overlapped. Finally, the quality of these images is often degraded due to the existence of noise, artifacts, or inhomogeneous background [12].

There is a considerable number of commercial software programs for analyzing 2-D gel images [9] including: PDQuest, DeCyder 2D, Melanie 7, and ImageMaster, as well as Delta2D and Progenesis Samespots. Despite their respective merits, each one has a different degree of success dealing with the existing challenges of 2-D gel images, and the best solution is far from being reached [9], [13]. First, they all require human intervention in order to specify mandatory input parameters. Second, they often present a number of drawbacks; they merge overlapping spots, split a single spot into more, fail to detect real spots, mistake artifacts for spots, and determine the boundary of spots without the desirable precision. It is worth mentioning that the aforementioned errors have an accumulative effect which inevitably modifies the protein expression levels thus leading to erroneous biological conclusions. Therefore, extensive manual editing, which is a time-consuming process, requiring 1 to 4 man-hours per gel on average, is needed to correct this multitude of errors [14]. It should be noted that human intervention not only limits the throughput but also brings the objectivity and reproducibility of results into question. Therefore, automating this part of the process is essential because: 1) it will allow expeditious high throughput analysis of the expression levels of thousands of proteins, and 2) it will lead to objective biological conclusions.

Manuscript received November 6, 2012; revised February 20, 2013; accepted April 11, 2013. Date of publication April 22, 2013; date of current version December 31, 2013. This work was supported in part by the European Union (European Social Fund—ESF) and in part by 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.

The authors are with the Department of Informatics and Telecommunications, University of Athens, Athens 15784, Greece (e-mail: ikostop@di.uoa.gr; eezacharia@gmail.com; dmaroulis@di.uoa.gr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JBHI.2013.2259208

Yoon *et al*. [15] proposed a method for the separation of overlapping spots as well as the enhancement of weak spots. Their method is based on the assumption that proteins have a Gaussian-distribution shape. Morris *et al*. [16] presented a spot detection and quantification method called "Pinnacle," while Li *et al*. [17] proposed a method for the quantification and statistical analysis of 2-D gel images called "RegStatGel." Both these methods process an average image of a properly aligned 2-D gel set. Pinnacle's spot detection is performed—on an average denoised image—by detecting "pinnacles" (i.e., local minima on the denoised average image) and combining them within a defined proximity. Although Pinnacle performs well in detecting the overlapping spots [16], it occasionally results in false positive spot detections [18]. RegStatGel's segmentation stage relies on the watershed algorithm [19] and faces difficulties in splitting overlapping spots [18]. Mylona *et al*. [20] proposed a method for the spot detection which is based on morphological operations. This method performs well in detecting the overlapping spots. However, it occasionally results in missed spots [20]. Dos Anjos *et al*. [11] proposed a spot segmentation and quantification method called "Scimo" which is also based on watersheds. This method can give a more realistic estimation of spots placed close to each other, as well as partially overlapping nonsaturated spots. However, it assumes that each basin of the watershed contains only one protein spot. In the case of major overlapping, the latter is not always valid. Recently, Savelonas *et al*. [10] proposed a segmentation method which exploits the properties of active contour formulation in order to deal with many of the challenges in 2-D gel images. Nevertheless, this method cannot segment overlapping spots.

In this paper, an original approach for the detection and segmentation of spots on a 2-D gel image is presented. The proposed approach is based on 2-D image histograms and on 3-D spots morphology. The centers of spots are determined using a clustering procedure, while spurious spots are removed using statistical measures. Earlier versions of our proposed approach presented in conferences [21]–[23] used neither the clustering procedure nor the statistical measures. These previous approaches have been evaluated on a limited number of spots and the results showed that they outperformed the compared software packages and techniques. However, they detect more spurious spots and miss actual protein spots. In this paper, the proposed approach has been evaluated using three datasets of 16-bit real and synthetic 2-D gel images, containing a total of ∼21 400 spots. To the best of our knowledge, this is the first attempt toward evaluating an approach for the spot detection and segmentation in 2-D gel images using such a large number of spots. Furthermore, the proposed approach has been compared with two commercial packages [Delta2D (http://www.decodon.com) and Melanie 7 (http://www.genebio.com)] as well as with two recently published methods [10], [11]. The results demonstrate that the proposed approach is very effective even when applied to images produced by different technology setups. These images contain: 1) spots of various shapes, sizes, and intensities, 2) several overlapping spots, as well as 3) noise, artifacts (streaks, speckles), and inhomogeneous background.

II. Proposed Approach

The proposed approach is divided into the following five successive steps: First, "regions of interest" (ROIs) containing mostly spots are specified based on the 2-D histogram of a 2-D gel image (A). Subsequently, the pixels that have a high probability of being spot-centers are determined—inside each ROI—based on the 3-D morphology of the protein spots (B). Afterward, spot surfaces are initially roughly (C) and then accurately (D) segmented. Finally, each spot surface is examined in order to establish whether the spot is real or spurious (E).

A. Specification of Regions of Interest on a 2-D Gel Image

Let $SR \subset \mathbb{N}^2$ be the set of ROIs on a 2-D gel image. The SR is defined by the following equation:

$$SR = \{R_r, r = 1, .., N\} \tag{1}$$

where N is the number of discrete regions, R_r of interest (ROIs) on the 2-D gel image (I). The SR set contains the pixels p of I that have high probability of forming spots regions. Since intensity values of spots vary, SR can be divided into two subsets: 1) a subset SP_H containing high-intensity pixels, and 2) a subset of pixels SP_L containing low-intensity pixels

$$SR = \{SP_H \cup SP_L\}. \tag{2}$$

The determination of the pixels being elements of SP_H or SP_L sets is accomplished by applying the 2-D Otsu thresholding technique [24]. It is worth mentioning that this thresholding technique does not require any parameters and is based on the 2-D histogram of the image, which—compared to the 1-D histogram—provides information about the spatial correlation between pixels on the image [25].

In our approach, the 2-D Otsu technique is applied twice. The SP_H and the SP_L sets are determined in the first and second iterations, respectively (see later).

1) Overview of the 2-D Histogram and 2-D Otsu Thresholding Technique: The 2-D histogram of an image I of $M \times N$ size is defined as

$$H(I(p), \bar{I}(p')) = \frac{o(I(p), \bar{I}(p'))}{M \times N} \tag{3}$$

where $\bar{I}$ denotes I after being smoothed with a 3×3 mean filter [26]. $I(p)$ denotes the intensity value of a pixel p in I, whereas $\bar{I}(p')$ is the intensity value of a pixel p'—which corresponds to pixel p—in $\bar{I}$. $o(I(p), \bar{I}(p'))$ stands for the occurrences of the intensity pairs $(I(p), \bar{I}(p'))$.

According to the 2-D Otsu thresholding technique, there is an optimal vector (S, T) which is automatically determined and which divides the 2-D histogram into four quadrilaterals (see Fig. 1) each having a specific attribute. The quadrilateral labeled as "2" contains the frequencies of the intensity pairs $(I(p), \bar{I}(p'))$ for which the respective pixels (p, p') are located in regions containing spots of high intensity. The quadrilateral labeled as "1" contains the frequencies of the pairs $(I(p), \bar{I}(p'))$ for which the respective pixels (p, p') are located in the background as well as in regions containing spots of low intensity. The quadrilaterals labeled as "3" and "4" contain the frequencies

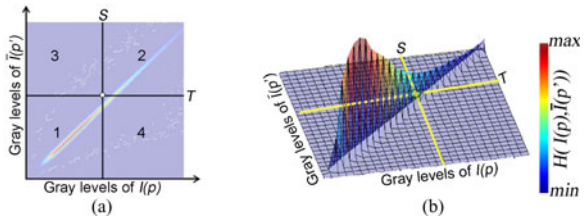

Fig. 1. Two-dimensional histogram of a typical 2-D gel image (a) in 2-D and (b) in 3-D representations.

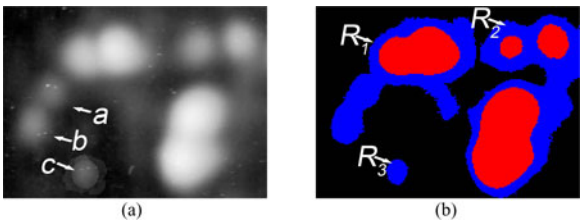

Fig. 2. Specification of ROI: (a) Part of a real 2-D gel image. (b) Three discrete regions of interest R_r, $r = 1, 2, 3$. SP_H and SP_L sets are colored in red and blue, respectively.

of the pairs $(I(p), \bar{I}(p'))$ for which the respective pixels (p, p') are located near edges or are noise.

2) Iterative 2-D Otsu Thresholding Technique: In the first iteration, the 2-D Otsu recursive thresholding technique is applied to the entire 2-D gel image (I). As mentioned earlier, the quadrilateral labeled as "2" corresponds to pixels having high probability of forming regions of high intensity spots. Therefore, the subset of pixels SP_H is defined as

$$SP_H = \{(p, p') : (I(p) \geq S) \wedge (\bar{I}(p') \geq T)\} \tag{4}$$

where (S,T) denotes a threshold vector which is automatically determined by applying the 2-D Otsu recursive thresholding technique to the entire 2-D gel image I.

In order to find the pixels having high probability of forming regions of low intensity spots and separate them from pixels of the background, 2-D Otsu recursive thresholding technique is subsequently applied to SP_H^C which is the complementary set of SP_H and it is defined as

$$SP_H^C = \{(p, p') : (p, p') \notin SP_H\}. \tag{5}$$

Likewise, the subset of pixels SP_L is defined as

$$SP_L = \{(p, p') : ((p, p') \in SP_H^C) \wedge (I(p) \geq S') \wedge (\bar{I}(p') \geq T')\} \tag{6}$$

where (S',T') denotes a threshold vector which is automatically determined using the 2-D Otsu recursive thresholding technique in the part of the image I which is formed by the pixels of the SP_H^C set.

An example of ROIs is depicted in Fig. 2. As one may observe, a significant portion of spot pixels—if not all [i.e., a, b, and c on Fig. 2(a)]—are not included in the SP_H set [colored in red, see Fig. 2(b)]. Instead, they are included in the SP_L set [colored in blue, see Fig. 2(b)]. Furthermore, each discrete ROI may contain one or more spots (for example, region R_1 contains four spots, while region R_3 contains a single spot).

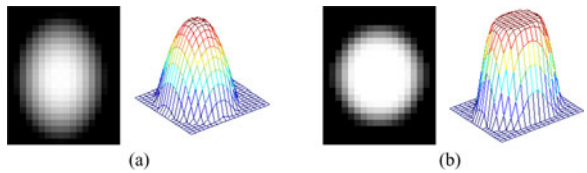

Fig. 3. Two-dimensional and 3-D representation of (a) a Gaussian spot and (b) a plateau spot.

B. Determination of Pixels Having High Probability of Being Spot-Centers

According to Bettens *et al.* [27], protein spot intensity peaks on spot central region and declines as distance from this region increases. In case the peak is thin, the 3-D morphological shape of the spot resembles a 3-D Gaussian function [see Fig. 3(a)], while when the peak is wide, the 3-D morphological shape of the spot resembles a plateau [see Fig. 3(b)]. Based on the aforementioned observation, it is self-explanatory that the spot-centers located in each distinct region R_r correspond to local intensity maxima of pixels contained within R_r. However, the opposite is not true as a 2-D gel image is often contaminated with noise as well as artifacts and contains inhomogeneous background. Consequently, the local maxima that have a high probability of being associated with spot-centers in each R_r are those which have the highest intensity value within a vicinity of local maxima. Therefore, local maxima in each region R_r are grouped together, and the highest local maximum of each group is considered to have a high probability of being associated with the actual spot-center.

In particular, let $G_r = (V, E)$ be a weighted graph, where V denotes the vertex-set and E denotes the edge-set of G_r. A vertex $v_m, m = 1, \ldots, |V|$ of the G_r graph represents a pixel located in a particular region R_r and corresponds to a local maximum of the median intensity values. In each vertex v_m, a weight equal to the median intensity value of the pixel that the vertex v_m represents is assigned. Each vertex v_m is connected with another one only if the Euclidean distance between the pixels they represent is less than T_d, where T_d is a constant.

A vertex v_m must be connected only to vertices of lesser weight in order to have high probability of being associated with a spot-center. If a vertex v_m is connected to a vertex of higher weight, then the v_m has low probability of being associated with a spot-center. If a vertex v_m is connected to l vertices $v_i, i = 1, \ldots, l$ of equal weights, then the v_m and the v_i vertices are considered as a single vertex for the purpose of comparing the weights of vertices.

Fig. 4 illustrates the region R_1 of Fig. 2. The local maxima having high and low probability of being associated with spot-centers are depicted with red and blue color, respectively. It should be highlighted that two of the local maxima (M_1, M_2)—that have high probability of being associated with spot-centers—do not correspond to actual spot-centers but to spurious spot-centers. In the last stage of our proposed approach, these local maxima are detected and removed.

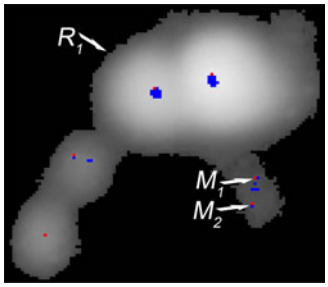

Fig. 4. Enlargement of the R_1 region of Fig. 2. The local maxima having high and low probability of being associated with spot-centers are illustrated with red and blue color, respectively.

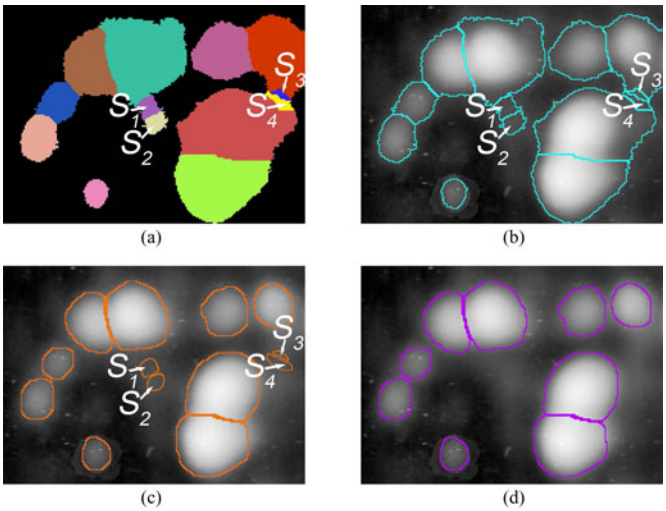

Fig. 5. (a) Roughly segmented spot surfaces and (b) their corresponding contours. (c) and (d) Precisely segmented spot surfaces. On (d) spurious spots have been detected and removed.

C. Rough Segmentation of Spot Surfaces

As mentioned earlier, each determined region R_r contains one or more local maxima that have a high probability of being associated with spot-centers. In this stage, the R_r region is segmented into subregions around the aforementioned local maxima.

In particular, each single vertex or group of equally weighted and connected vertices—each having high probability of being associated with the spot-center—forms a distinct cluster C_k, where $k \subset \mathbb{N}$. The proposed algorithm proceeds by joining each pixel p_i^* of median intensity value i^* to the cluster with which it is mostly connected, for all median intensities i^* between $i^*_{\max}$ and $i^*_{\min}$. Therefore, for each p_i^* pixel a majority voting criterion is applied amongst its 3×3 adjacent pixels. If none of its adjacent pixels belongs to a cluster, then the algorithm proceeds by joining the rest of the pixels of median intensity i^* before joining the p_i^* pixel to a cluster.

The subregions in which the regions R_r have been segmented are illustrated in Fig. 5(a) with different colors. The boundaries of these roughly segmented spot surfaces are depicted in Fig. 5(b). As one can observe, each of the R_1 and R_2—each containing only four protein spots—has been segmented into six subregions. The four spurious spots—denoted as S_1, S_2, S_3, and S_4 in Fig. 5(a)—are detected and removed in the last stage of the proposed approach.

D. Precise Segmentation of Spot Surfaces

Let R_{rj} be the j subregion of the R_r region. The spot surface located inside the R_{rj} is segmented as follows. First, the R_{rj} subregion is expanded by adding pixels p around its perimeter so that the pixels p are not contained simultaneously in another adjacent subregion R_{rk}, where $k \neq j$. In this respect, the dilation operator in the R_{rj} region is applied using a disk D of radius r_{D} as a structuring element [28]. In particular, the expanded region R^E_{rj} of R_{rj} is defined as

$$R^E_{rj} = \{p : (p \in (R_{rj} \oplus D)) \wedge (p \notin (R_{rk} \oplus D))\} \quad (7)$$

where $R_{rj} \oplus D$ stands for the dilated region R_{rj} using the structuring element D.

Subsequently, the iterative (optimal) thresholding technique [29] is applied in the gradient intensity values of pixels located in R^E_{rj} in order to obtain the optimal contour of the spots. The precise spot surface is defined as the area, which is inside the elliptical shape determined by the pixels having the highest gradient intensity values. For example, Fig. 5(c) illustrates the contour of the spot surfaces precisely segmented.

E. Detection and Elimination of Spurious Spots

As mentioned earlier, a number of spot surfaces may not contain actual protein spots but spurious ones. These spot surfaces are likely to contain either many local maxima (in the case of noise and artifacts) or are likely to be highly homogeneous (in the case of background). Based on this observation, spurious spots are those which fulfill one of the following two conditions.

1) If the percentage of local maxima inside a spot surface is higher than a maximum acceptable threshold T_m, then this spot surface contains a spurious spot.
2) If the coefficient of the variation CV [30]–[32] in the intensity values inside a spot surface is lower than a minimum acceptable value T_{cv} then this spot surface contains a spurious spot. Indeed, if the CV is lower than T_{cv} it means that the dispersion of the intensity distribution is small, thus the spot surface is in reality a background region. Fig. 5(d) illustrates the segmentation result after the spurious spots removal.

III. Results and Discussion

Several experiments were conducted in order to evaluate the performance of the proposed approach and compare it with two established commercial packages (Melanie 7 and Delta2D) as well as with two recently published methods (Savelonas *et al.* [10] and Scimo [11]). In this respect, we used 16-bit images obtained from three different datasets, containing a total of ~21 400 spots, while existing techniques have been tested on significantly lower number of spots for their evaluation (e.g., the method of dos Anjos *et al.* [11] has been tested on ~1000 spots). It is worth mentioning that the proposed approach supports images of any bit depth. However, the 16-bit images were prior

candidates, since they are characterized by high resolution which enables low-intensity spots to be more easily distinguished from the background, in order to enhance their visualization. A user-friendly software implementation of the proposed approach, named "ANGELI" (Analyzing 2-D Gel Electrophoresis Images), is available online at http://rtsimage.di.uoa.gr/ANGELI.

The first two datasets (D_1, D_2) consist of real 2-D gel images containing a total of $\sim$10 200 and $\sim$1000 spots, respectively. D_1 has been provided through the courtesy of the Biomedical Research Foundation of the Academy of Athens (http://www.bioacademy.gr), whereas D_2 is the dataset that has been used by dos Anjos *et al*. [11] in order to evaluate their algorithm—named "Scimo"—as well as to compare it with two other state-of-the-art programs. The third dataset D_3 consists of synthetic 2-D gel images which have been produced by our group. These images contain a total of $\sim$10 200 spots. Each spot was produced by using the 2-D Gaussian flat top function. In order to create synthetic images that look similar to the real ones, the spots were overlaid onto backgrounds extracted from real 2-D gel images. It should be noted that the three datasets contain single, as well as overlapping spots of various intensities, sizes, and shapes. The spots are surrounded by inhomogeneous noisy background that also contains artifacts.

The parameters of the proposed approach were adjusted once, and they remained constant during all experiments performed on the three datasets. Thus, the whole experimental procedure on the synthetic and real 2-D gel images took place without any human intervention. In particular, the constant T_d of 8 was adopted as the maximum acceptable distance between two connected vertices. A disk D of radius 4 was adopted for the dilation ($r_{\mathrm{D}} = 4$). The maximum acceptable value T_m of 2.5% and the minimum acceptable value T_{CV} of 0.001 were chosen in order to distinguish the real spots from the spurious spots. It should be highlighted that although the parameters of the proposed approach were adjusted once—not in accordance with each particular image—the parameters of the other four existing software programs and techniques were adjusted during the evaluation—by expert biologists—according to each separate image. In this way, the results of the proposed approach using the default parameters were compared on purpose with the optimal results of the other commercial programs and techniques.

In order to statistically analyze the detection results of the proposed approach with the aforementioned commercial packages and techniques, the D_1 and D_2 datasets of real 2-D gel images were used. The ground truth for these datasets was provided by expert biologists of the Biomedical Research Foundation of the Academy of Athens (http://www.bioacademy.gr), who manually determined the locations of protein spots by drawing a cross inside each unique spot-region. Each image had three ground truth replicates and its final ground truth was determined by majority vote from these replicates.

The detection results were quantitatively evaluated using sensitivity S, precision P as well as their weighted harmonic mean (F-measure) defined as

$$S = \frac{TP}{TP + FN}, \quad P = \frac{TP}{TP + FP} \tag{8}$$

TABLE I
COMPARISON OF DETECTION RESULTS ON REAL 2-D GEL IMAGES (D_1 AND D_2 DATASETS)

	%	P	S	F
D_1	Melanie 7	85.2±2.8	94.1±2.4	89.4±2.4
	Delta2D	75.7±15.9	78.1±6.2	76.9±10.6
	Scimo	91.9±3.7	76.4±3.6	83.4±3.3
	Savelonas *et al.*	92.5±2.5	66.3±7.1	77.2±5.5
	Proposed approach	**92.8**±3.6	**94.0**±1.9	**93.4**±2.2
D_2	Melanie 7	81.7±14.1	94.8±1.1	87.4±8.7
	Delta2D	48.4±17.8	88.2±4.2	59.2±16.5
	Scimo	75.0±18.5	86.5±2.4	78.5±12.6
	Savelonas *et al.*	94.5±9.0	65.2±8.1	77.2±4.4
	Proposed approach	**99.3**±0.2	**95.3**±4.5	**97.2**±2.4

$$F\text{-measure} = 2 \times \frac{S \times P}{S + P} \tag{9}$$

where FN (false negatives), TP (true positives), and FP (false positives) denote, respectively, the number of spots that are missed, correctly detected, and falsely detected by a specific method.

Sensitivity and precision values are both important, complementary measures on the detection performance. However, F-measure value is a more reliable measure than just sensitivity and precision, as it takes into account both the number of detected protein spots as well as the number of spurious spots.

Table I presents a comparison of detection results between the proposed approach and four other published methods and commercial packages using the D_1 and D_2 datasets. Based on the results, it is evident that the proposed approach: 1) detects protein spots on single gels effectively, in both datasets and 2) is more successful than the other four software programs and techniques in detecting spots. In particular, it achieves values greater than 92% in each measure (precision, sensitivity, and F-measure) and more specifically, for the F-measure it achieves the highest value against the compared software packages and techniques (93.4% on the D_1 dataset and 97.2% on the D_2 dataset). These high values of F-measure mean that: 1) the proposed approach detected almost all the real spots in the 2-D gel images (sensitivity value is also very high), even though the images were produced by means of different technologies, and 2) it detected a negligibly small number of spurious spots (precision value is very high too). The latter remark is significant since both high precision and high sensitivity minimize the effort and time needed for validation and correction of the results by expert biologists. Contrary to the proposed approach, the other four techniques achieve a lower value in precision, sensitivity, or both. According to the aforementioned remark, all these techniques have detected a higher percentage of spurious spots or have overlooked more spots than the proposed approach. For example, in the D_1 dataset, Melanie 7 has overlooked less spots than the proposed approach (0.1% higher sensitivity value than the proposed approach) but has simultaneously detected a much higher percentage of spurious spots (7.6% lower precision value). It is evident that the lower precision value of Melanie 7 ($\sim$85%) compared to the proposed approach's ($\sim$93%) demands more time and effort for the validation and correction of the results by expert biologists.

TABLE II
COMPARISON OF SEGMENTATION RESULTS ON SYNTHETIC 2-D GEL IMAGES (D_3 DATASET)

%	P	S	F	E
Melanie 7	93.7 ±8.5	82.5±6.8	87.8±5.4	14.7±4.7
Delta2D	61.8±11.6	97.3±7.4	75.6±9.3	32.6±9.0
Scimo	94.9±5.5	82.3±7.7	88.2±4.9	17.9±5.1
Savelonas *et al.*	95.9±10.2	61.5±18.1	74.9±13.9	34.9±13.7
Proposed approach	**92.5**±6.7	**94.2**±6.2	**93.3**±5.9	**8.6**±2.9

In order to statistically analyze the segmentation results of the proposed approach against the aforementioned packages and techniques, the D_3 dataset of synthetic 2-D gel images was used, for which the exact boundary of each individual spot was known. The segmentation results were quantitatively evaluated using sensitivity, precision, F-measure as well as the normalized error E between the real spot volume—based on the ground truth—and the estimated spot volume by the respective segmentation method. The normalized error E has already been used in the statistical analysis performed by [33].

Table II presents a comparison of the segmentation results for the D_3 dataset. Based on the results, it is evident that the proposed approach segments the spots in a more effective manner than the other four software programs and techniques. In particular, its F-measure has the highest value (93.3%), while both the precision and sensitivity values are very high too. Melanie 7, Scimo, and Savelonas *et al*. methods achieve a slightly higher precision value, yet a significantly lower sensitivity value than the proposed approach. Their higher precision value corresponds to less background-pixels being included in the spot-region. On the contrary, their lower sensitivity value corresponds to more spot-pixels being excluded from the spot-region, compared to the proposed approach. Based on this remark, precision and sensitivity values are complementary measures. Therefore, F-measure is a more reliable measure than just sensitivity or precision, as it takes into account both the number of background-pixels included in the spot region and the number of spot-pixels excluded from it. Furthermore, the proposed approach's error margin is limited to 8.6% while the second best performing software program (Melanie 7) bears an error margin of 14.7%. In other words, the error margin of Melanie 7 is >50% greater than the proposed approach.

Paired student's t-test was also performed comparing the segmentation errors of all spots between our proposed approach and the preceding ones (Melanie 7, Delta2D, Scimo, and Savelonas *et al*. method). These tests showed that there was a significant difference ($p < 10^{-6}$) between the error of the proposed approach and the errors of Melanie 7, Delta2D, Scimo, and Savelonas *et al*. method.

Figs. 6 and 7 illustrate the segmentation results produced by the proposed approach on a real and a synthetic 2-D gel image obtained from D_1 and D_3 dataset, respectively. Figs. 8, 9, and 10 illustrate three 2-D gel subimages each one obtained from a different dataset (D_1, D_2, and D_3, respectively) and the segmentation results produced by Melanie 7, Delta2D, Scimo, and Savelonas *et al*. and the proposed approach. In the aforementioned images, the locations of the spots—according to ground

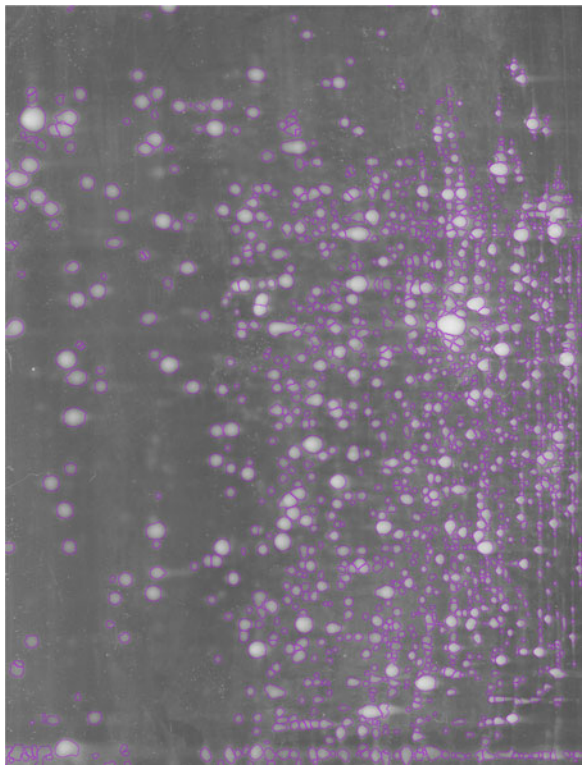

Fig. 6. Segmentation result of a real 2-D gel image from D_1 dataset using the proposed approach.

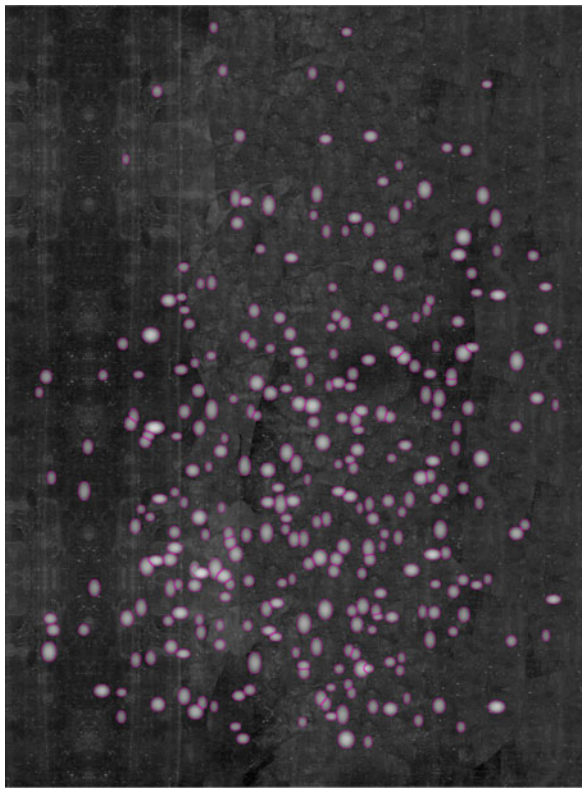

Fig. 7. Segmentation result of a synthetic 2-D gel image from D_3 dataset using the proposed approach.

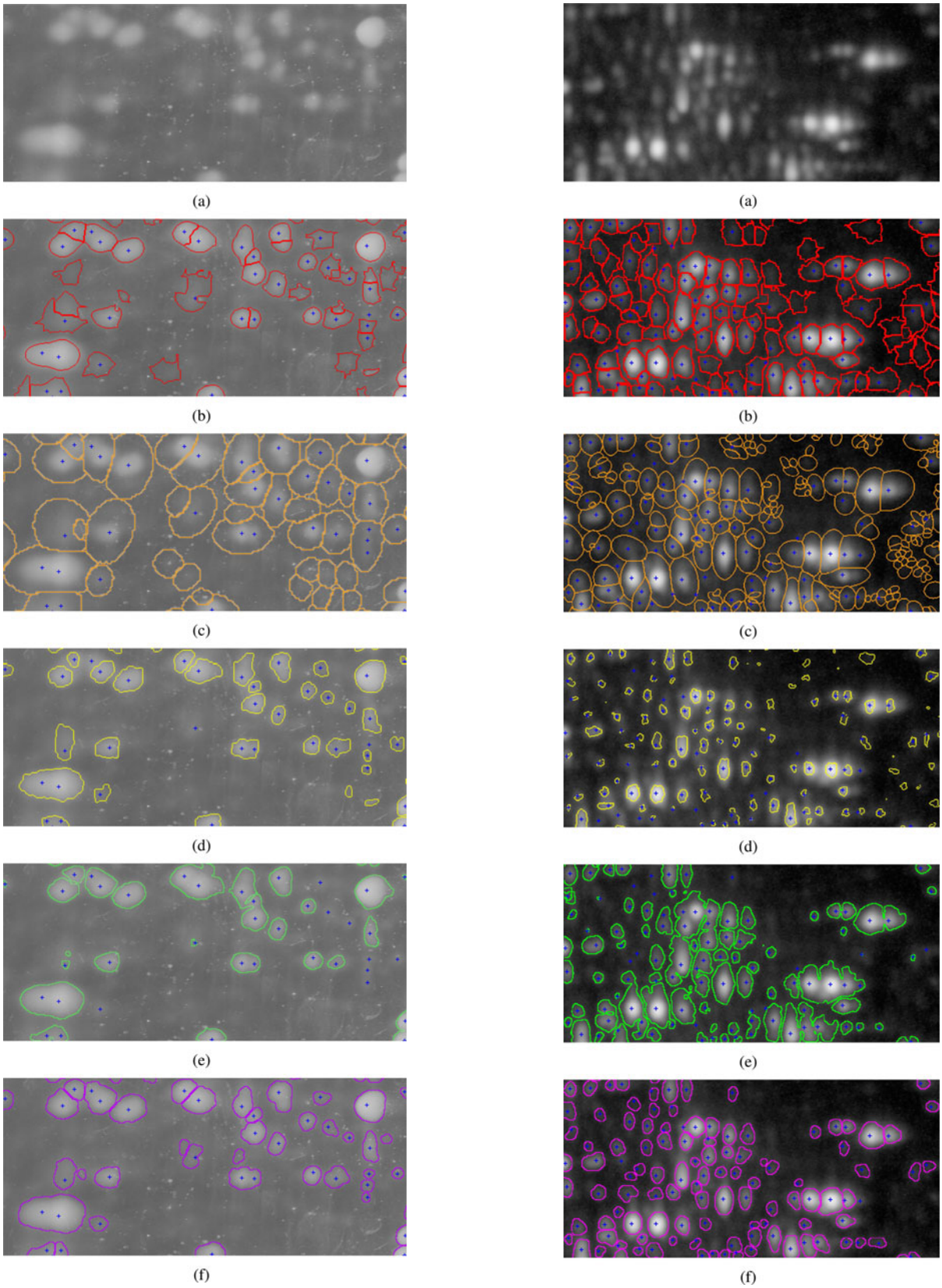

Fig. 8. Real subimage (a) from the D_1 dataset with its segmentation results using (b) Melanie 7, (c) Delta2D, (d) Scimo, (e) Savelonas *et al.*, and (f) the proposed approach.

Fig. 9. Real subimage (a) from the D_2 dataset with its segmentation results using (b) Melanie 7, (c) Delta2D, (d) Scimo, (e) Savelonas *et al.*, and (f) the proposed approach.

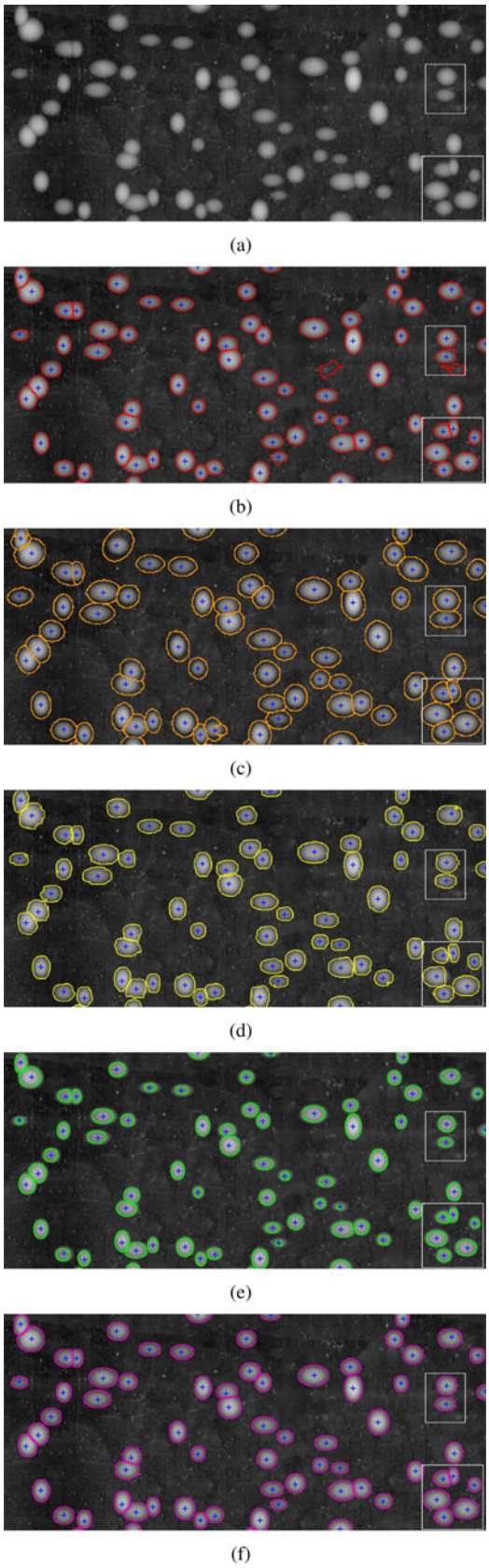

Fig. 10. Synthetic subimage (a) from the D_3 dataset with its segmentation results using (b) Melanie 7, (c) Delta2D, (d) Scimo, (e) Savelonas *et al*., and (f) the proposed approach.

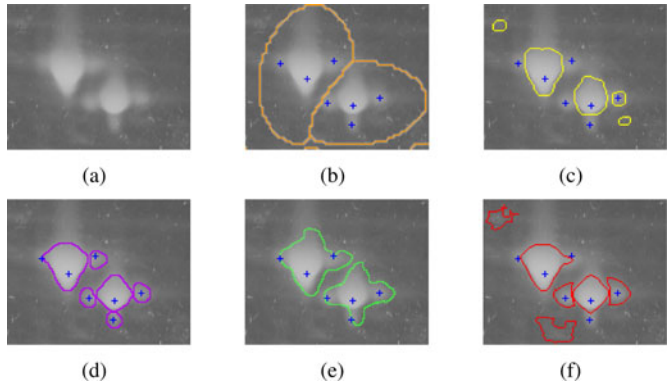

Fig. 11. Real subimage (a) from the D_1 dataset with its segmentation results using (b) Melanie 7, (c) Delta2D, (d) Scimo, (e) Savelonas *et al*., and (f) the proposed approach.

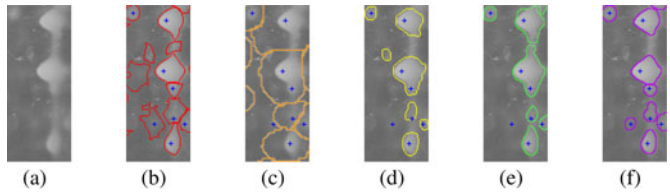

Fig. 12. Real subimage (a) from the D_1 dataset containing a streak with spots as well as its segmentation results using (b) Melanie 7, (c) Delta2D, (d) Scimo, (e) Savelonas *et al*., and (f) the proposed approach.

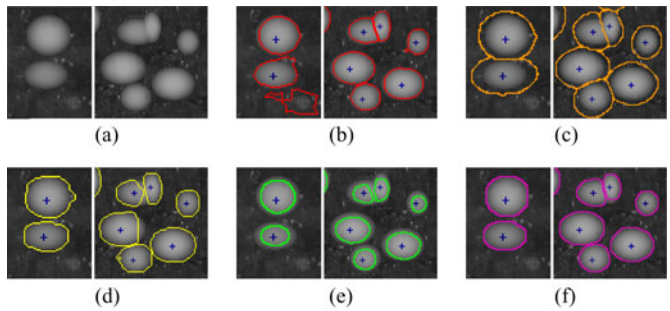

Fig. 13. Enlargement of the two white oblongs of Fig. 8 with their segmentation results using (b) Melanie 7, (c) Delta2D, (d) Scimo, (e) Savelonas *et al*., and (f) the proposed approach.

truth—are illustrated with blue crosses. As mentioned earlier, the ground truth of real images has been provided by expert biologists.

Based on the results, it is evident that the proposed approach has very effectively detected and segmented the protein spots. For instance, in Fig. 9 it has detected 100 protein spots out of 102, while Melanie 7, Delta2D, Scimo, and Savelonas *et al*. method have detected 90, 88, 91, and 67 real spots, respectively. Furthermore, in the same image the proposed approach found 1 spurious spot, while Melanie 7, Delta2D, Scimo, and Savelonas *et al*. method found 18, 83, 21, and 3 spurious spots, respectively. Last but not least, the proposed approach has segmented the protein spots more accurately in contrary to the other methods which either include background within the spot area or exclude a part of the spot area.

This is illustrated in Figs. 11, 12, and 13 which depict the segmentation results of four 2-D gel subimages (two real and

two synthetic). According to the ground truth of these images (blue crosses), the proposed approach has effectively segmented the protein spots without detecting any spurious spot, even if the subimage contains a streak (see Fig. 12). However, the results obtained by the other software programs and techniques need manual correction. Moreover, the spot boundaries generated by the proposed approach are more plausible than those generated by the software packages and techniques.

IV. Conclusion

Spot detection and segmentation in 2-D gel images are undoubtedly two challenging stages of the proteomic analysis sequence. In this paper, an original approach for the detection and segmentation of 2-D gel spots is presented. The proposed approach is very effective under the following adverse conditions: 1) the presence of various spot-shapes, sizes, and intensities, 2) the presence of overlapping spots, as well as 3) the presence of artifacts (streaks, speckles), noise, and inhomogeneous background. The experimental results over synthetic and real images confirm the validity of our method, as well as its accuracy and effectiveness. Additionally, its default parameters worked well regardless of the nature of these images. Overall, the results suggest that the proposed approach is a promising alternative to the state-of-the-art published methods.

Acknowledgment

The authors would like to extend their gratitude to the Biomedical Research Foundation of the Academy of Athens and in particular to Dr. S. Kossida and Dr. A. Vlahou for providing the real 2-D gel images (D_1 dataset), the ground truth of both real D_1 and D_2 datasets, and the segmentation results obtained by Melanie 7 software program and Scimo (for D_1 and D_3 datasets). Furthermore, the authors would like to thank expert biologist Dr. M. Aivaliotis for providing the segmentation results obtained by Delta2D software package. Finally, we would like to thank the authors of Scimo [11] for providing the D_2 dataset and its segmentation results, as well as for their constructive comments on their technique.

References

[1] G. Klaiman, T. Petzke, J. Hammond, and A. LeBlanc, "Targets of caspase-6 activity in human neurons and alzheimer disease," *Molecular Cell. Proteomics*, vol. 7, no. 8, pp. 1541–1555, 2008.

[2] K. Kultima, B. Scholz, H. Alm, K. Sköld, M. Svensson, A. Crossman, E. Bezard, P. Andrén, and I. Lönnstedt, "Normalization and expression changes in predefined sets of proteins using 2D gel electrophoresis: A proteomic study of l-DOPA induced dyskinesia in an animal model of Parkinson's disease using DIGE," *BMC Bioinf.*, vol. 7, no. 1, pp. 475–501, 2006.

[3] J. Lee, J. Han, G. Altwerger, and E. Kohn, "Proteomics and biomarkers in clinical trials for drug development," *J. Proteomics*, vol. 18, pp. 2632–2641, 2011.

[4] M. Natale, D. Bonino, P. Consoli, T. Alberio, R. Ravid, M. Fasano, and E. Bucci, "A meta-analysis of two-dimensional electrophoresis pattern of the parkinson's disease-related protein DJ-1," *Bioinf.*, vol. 26, no. 7, pp. 946–952, 2010.

[5] A. Dowsey, M. Dunn, and G. Yang, "The role of bioinformatics in two-dimensional gel electrophoresis," *Proteomics*, vol. 3, no. 8, pp. 1567–1596, 2003.

[6] J. Lopez, "Two-dimensional electrophoresis in proteome expression analysis," *J. Chromatograph. B*, vol. 849, no. 1–2, pp. 190–202, 2007.

[7] W. Van Belle, N. Ånensen, I. Haaland, Ø. Bruserud, K. Høgda, and B. Gjertsen, "Correlation analysis of two-dimensional gel electrophoretic protein patterns and biological variables," *BMC Bioinf.*, vol. 7, no. 1, 2006.

[8] T. Aittokallio, J. Salmi, T. Nyman, and O. Nevalainen, "Geometrical distortions in two-dimensional gels: Applicable correction methods," *J. Chromatograph. B, Anal. Technol. Biomed. Life Sci.*, vol. 815, no. 1–2, pp. 25–37, 2005.

[9] S. Magdeldin, Y. Zhang, B. Xu, Y. Yoshida, and T. Yamamoto, "Two-dimensional polyacrylamide gel electrophoresis—A practical perspective," *Gel Electrophor.—Principles Basics*, pp. 91–116, 2012.

[10] M. Savelonas, E. Mylona, and D. Maroulis, "Unsupervised 2D gel electrophoresis image segmentation based on active contours," *Pattern Recog.*, vol. 45, no. 2, pp. 720–731, 2012.

[11] A. dos Anjos, A. Møller, B. Ersbøll, C. Finnie, and H. Shahbazkia, "New approach for segmentation and quantification of two-dimensional gel electrophoresis images," *Bioinformatics*, vol. 27, no. 3, pp. 368–375, 2011.

[12] A. Görg, W. Weiss, and M. Dunn, "Current two-dimensional electrophoresis technology for proteomics," *Proteomics*, vol. 4, no. 12, pp. 3665–3685, 2004.

[13] B. Clark and H. Gutstein, "The myth of automated, high-throughput two-dimensional gel analysis," *Proteomics*, vol. 8, no. 6, pp. 1197–1203, 2008.

[14] P. Cutler, G. Heald, I. White, and J. Ruan, "A novel approach to spot detection for two-dimensional gel electrophoresis images using pixel value collection," *Proteomics*, vol. 3, no. 4, pp. 392–401, 2003.

[15] J. W. Yoon, S. J. Godsill, C. Kang, and T.-S. Kim, "Bayesian inference for 2D gel electrophoresis image analysis," in *Proc. 1st Int. Conf. Bioinf. Res. Develop*, 2007, pp. 343–356.

[16] J. Morris, B. Clark, and H. Gutstein, "Pinnacle: A fast, automatic and accurate method for detecting and quantifying protein spots in 2-dimensional gel electrophoresis data," *Bioinformatics*, vol. 24, no. 4, pp. 529–536, 2008.

[17] F. Li, F. Seillier-Moiseiwitsch *et al.*, "Differential analysis of 2D gel images.," *Methods Enzymol.*, vol. 487, pp. 595–605, 2011.

[18] Y. Wu and L. Zhang, "Comparison of two academic software packages for analyzing two-dimensional gel images," *J. Bioinf. Comput. Biol.*, vol. 9, no. 6, pp. 775–794, 2011.

[19] L. Vincent and P. Soille, "Watersheds in digital spaces: An efficient algorithm based on immersion simulations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 6, pp. 583–598, Jun. 1991.

[20] E. Mylona, M. Savelonas, D. Maroulis, and S. Kossida, "A computer-based technique for automated spot detection in proteomics images," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 4, pp. 661–667, Jul. 2011.

[21] E. Zacharia, E. Kostopoulou, D. Maroulis, and S. Kossida, "A spot segmentation approach for 2D gel electrophoresis images based on 2D histograms," in *Proc. IEEE 20th Int. Conf. Pattern Recog.*, Aug. 2010, pp. 2540–2543.

[22] E. Kostopoulou, E. Zacharia, and D. Maroulis, "Detection and segmentation in 2D gel electrophoresis images," in *Proc. IEEE 17th Int. Conf. Digital Signal Process.*, Jul. 2011, pp. 1–6.

[23] E. Kostopoulou, E. Zacharia, and D. Maroulis, "Accurate segmentation of 2D-page images," in *Proc. IEEE 20th Eur. Conf. Signal Process.*, 2012, pp. 1–5.

[24] J. Zhang and J. Hu, "Image segmentation based on 2D Otsu method with histogram analysis," in *Proc. IEEE Int. Conf. Comput. Sci. Softw. Eng.*, Dec. 2008, vol. 6, pp. 105–108.

[25] J. Gong, L. Li, and W. Chen, "Fast recursive algorithms for two-dimensional thresholding," *Pattern Recog.*, vol. 31, no. 3, pp. 295–300, 1998.

[26] J. C. Russ, *Image Processing Handbook*, 4th ed. Boca Raton, FL, USA: CRC Press, 2002.

[27] E. Bettens, P. Scheunders, D. Van Dyck, L. Moens, and P. Van Osta, "Computer analysis of two-dimensional electrophoresis gels: A new segmentation and modeling algorithm," *Electrophoresis*, vol. 18, no. 5, pp. 792–798, 1997.

[28] R. Gonzalez and R. Woods, *Digital Image Processing*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2002.

[29] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis, and Machine Vision*. Florence, KY, USA: Cengage-Engineering, 2007.

[30] M. Katajamaa and M. Orešič, "Processing methods for differential analysis of lc/ms profile data," *BMC Bioinf.*, vol. 6, no. 1, pp. 179–190, 2005.

[31] F. Millenaar, J. Okyere, S. May, M. Van Zanten, L. Voesenek, and A. Peeters, "How to decide? Different methods of calculating gene expression from short oligonucleotide array data will give different results," *BMC Bioinf.*, vol. 7, no. 1, pp. 137–152, 2006.

[32] I. Miller, J. Freund, and R. Johnson, *Probability and Statistics for Engineers*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1965, vol. 4.
[33] P. Peer, L. Corzo *et al*., "Local pixel value collection algorithm for spot segmentation in two-dimensional gel electrophoresis research," *Comparat. Funct. Genom*., vol. 2007, art. ID 89596, 2007.

Eirini Kostopoulou received the B.Sc. degree in computer science from the Technological Educational Institute of Athens, Athens, Greece, in 2007, and the M.Sc. degree in bioinformatics (Hons.) after attending the "Information Technologies in Medicine and Biology" postgraduate program at the University of Athens, Athens, Greece, where she is currently working toward the Ph.D. degree in image analysis.

She was awarded the State Scholarships Foundation (IKY) scholarship for 2 years given to the student with the best grades. For her postgraduate studies, she received a four-year scholarship after participation in IKY exams (1st rank). She has coauthored five research articles on biomedical image analysis. Her research interests include biomedical image analysis, segmentation, and bioinformatics.

Eleni Zacharia received the B.Sc. degree in informatics and telecommunications, the M.Sc. in signal processing for communications and multimedia with the highest degree among her peers, and the Ph.D. degree in computer science from the Department of Informatics and Telecommunications, University of Athens, Athens, Greece, in 2004, 2006, and 2009, respectively. For Ph.D. research, she received a scholarship from the Greek General Secretariat of Research and Technology (25%) and the European Social Fund (75%).

In 2011–2012, she was awarded a Postdoctoral Fellowship from the Keck Center Computational Cancer Biology Training Program of the Gulf Coast Consortia, Houston, TX, USA. She was a Postdoctoral Researcher at the Computational Biomedicine Lab, Department of Computer Science as well as the Center for Nuclear Receptors and Cell Signaling, both at University of Houston, Houston, TX, USA. She is currently a Research Fellow in the Department of Informatics and Telecommunications, University of Athens. Her research interests include the areas of biomedical image analysis and pattern recognition.

Dimitris Maroulis (M'02) received the B.Sc. degree in physics, the M.Sc. degree (Hons.) in radioelectricity and in cybernetics, and the Ph.D. degree (Hons.) in computer science, all from the University of Athens, Athens, Greece.

He served as a Research Fellow for 3 years at the Space Research Department (DESPA) of Meudon Observatory, Paris, France, and afterward he collaborated for more than 10 years with the same department. He has also served in various academic positions in the Departments of Physics and Informatics of the University of Athens where he is currently a Professor in the Department of Informatics and Telecommunications and the Leader of the Real Time Systems and Image Analysis Lab. He has more than 20 years of experience in the areas of data acquisition and real-time systems, and more than 15 years of experience in the area of image/signal analysis and processing. He has also been collaborating with many Greek and European hospitals and health centers for more than 15 years in the field of biomedical informatics. He has been actively involved in more than 15 European and National R&D projects and has been the Project Leader of five of them, all in the areas of image/signal analysis and real-time systems. He has published more than 150 research papers and book chapters, and there are currently more than 1000 citations that refer to his published work. His research interests include data acquisition and real-time systems, pattern recognition, image/signal processing and analysis, with applications on biomedical systems, and bioinformatics.