Image processing methods and algorithms for accurate protein spot detection in 2-dimensional gel electrophoresis (2DGE)

Panagiotis Tsakanikas[1]

National and Kapodistrian University of Athens
Department of Informatics and Telecommunications
tsakanik@di.uoa.gr

Abstract. The main goal of this dissertation is the development of methods to improve the accuracy and efficiency of the protein spot detection and quantification on 2DGE images. Image analysis is still considered as the bottleneck of the differential expression proteomics analysis workflow, due to the large variability in the protein spots' expression profiles where a lot of manual user work is needed in order to achieve acceptable results. The contributions of this dissertation are apparent to all the stages of a 2DGE image analysis: (i) development of a new and specialized method for image denoising based on multiresolution analysis and the Contourlet Transform which was evaluated with synthetic and real images and shown that it outperforms the existing denoising approaches in terms of noise suppression and optimizing the subsequent analysis results; (ii) development of a novel approach for delineating 2DGE image areas which -with high probability- include protein spots, based on Active Contours. The developed method has been evaluated using a large pool of synthetic and real gel images and shown that the extracted ROIs include the large majority of the true spots while it functions in an fully automatic way; (iii) novel hierarchical approach to protein spot segmentation based on machine learning techniques and Gaussian mixture models. The developed approach is applied on each previously extracted Region of Interest (ROI), aiming at removing local background and streaks, while estimating the number, location and borders of proteins spots contained. After an exhaustive evaluation the developed methodology proved to be more accurate and efficient than competing methods while it is grounded on the physical properties of protein spots and overlapping spots formation.

Keywords: Proteomics, two-dimensional gel electrophoresis, denoising, image segmentation, spot detection, spot quantification, spot modeling.

1 Introduction

During the last decade, the life and computer science communities are striving to build models in order to develop a global understanding of the living cell. This effort is deeply influenced by the development of the "omics" technologies (genomics, transcriptomics, proteomics, metabolomics, etc), which aim at establishing a holistic view on biological systems. *Proteomics* is the large-scale study of proteins, and in particularly of their structures and functions [1,2]. Proteins are vital parts of living cells, as they are the main components of the physiological metabolic pathways. The proteome is the entire set of proteins [3] expressed by an organism or system (including the modifications made to a particular set of proteins). This varies with time and depends on the stresses that a cell or organism undergoes. So, proteomics is the study of the time varying proteome using the technologies of large-scale protein separation and identification. In other words it is the study of proteins, how they are modified, when and where they are expressed, how they are involved in signaling and metabolic pathways and how they interact with each other. Current research in proteomics requires that proteins in a biological sample be effectively resolved. To achieve this goal, proteins need to be separated first. This separation can be performed using two-

[1] Dissertation Advisor: Elias S. Manolakos, Associate Professor

dimensional gel electrophoresis (2DGE) and gives rise to protein spots of irregular shapes and sizes on the gel. Once proteins are separated and quantified, they can be identified. For this purpose, individual spots are cut out of the gel and cleaved into peptides with proteolytic enzymes. These peptides can then be identified using mass spectrometry methods combined with database search.

Two-dimensional gel electrophoresis is a powerful and widely used method for the analysis of complex protein mixtures extracted from cells, tissues, or other biological samples. This technique is used to separate the proteins in two steps according to two independent properties: isoelectric focusing (IEF), which separates proteins according to their isoelectric points (pI), and SDS-polyacrylamide gel electrophoresis (SDS-PAGE), which separates proteins according to their molecular weights (MW). In this way, complex mixtures, consisted of thousands of different proteins, can be resolved and the relative amount of each protein can then be determined. The resulting gel is then digitized and an image analysis pipeline is applied in order to find protein spot specific properties (i.e. location, quantity, etc). In general, a typical 2DGE image processing pipeline consists of the following stages [4]:

1. Image pre-processing (noise suppression, artifacts removal, streak removal and background correction.
2. Spot segmentation (spot detection). Delineate each individual spot area, outputting a list of spot centers, intensities and geometric features. The spot detection operations pipeline in most cases is:
    1. Detect the centers of as many spots as possible.
    2. Segment the gel into regions, each containing one of these spots.
3. Modeling and quantification (spot volume estimation). Model each extracted spot region by a parametric spot model in order to extract a characteristic vector for each spot for further data analysis, and to detect and separate co-migrated spots.
4. Corresponding protein spot matching across gels of different samples.
5. Identification of differential expression (using statistical methods).

1.1 Related Work and Current Limitations

Although the resolution of 2DGE seems impressive, it is still not sufficient compared to the enormous diversity of cellular proteins, and co-migrating proteins in the same spot are not uncommon [5]. Neighboring spots can obscure protein spot centers in these so-called complex regions, and their saturated nature can make the resolution of each individual protein impossible. Furthermore, spots tend to have symmetric diffusion in the pI dimension but often severe tails in the Mr dimension (streaks). This diffusion depends on the protein concentration, which is why streaks and smears occur with certain proteins. So, the “faint” protein spots require an expert eye to discriminate them from noise, so an efficient noise suppression method would be extremely useful. Also, the intensity of the background can vary across the image. Finally, in most cases there are incompletely separated (overlapping) spots (less-defined and/or separated) and several complex protein spot areas. Inefficiency due to the above limitations leads also to unmatched/undetected spots (leading to missing values), mismatched spots, and errors in quantification (several distinct spots may be erroneously detected as a single spot by the software and/or parts of a spot may be excluded from quantification). All of the above constitute some of the big challenges for the automation of spot detection in the bioinformatics pipeline. Throughout the years a lot of commercial and non-commercial methods for image analysis of gel images have been developed.

Most commercially available 2DGE image analysis tools use conventional spatial filters [7,11] to combat with noise, mainly due to the fact that they are conceptually simple and computationally efficient. However, spatial filtering introduces severe distortions at protein spot borders and alters considerably the intensity values of internal spot pixels [6]. To address spatial filtering limitations, multiresolution space-frequency domain techniques based on wavelets have been proposed [6]. It has been shown that the Wavelet Transform (WT) outperforms spatial filtering both in terms of signal-to-noise ratio (SNR) performance and in terms of the resulting visual image quality [6]. Despite its several advantages, the Wavelet Transform has also some notable limitations [12]. The most relevant to 2DGE image denoising is its limited ability in capturing directional information, as needed to adequately represent the smoothness along spot boundaries.

Another pre-processing operation is the background subtraction used in order to eliminate meaningless changes in the gel background intensity level. A simple approach is to obtain the lightest and darkest point in the background and replace the whole background with the average intensity. Tyson and Haralick [8] find the local minima in the image, representing background depressions, and interpolate the background between these minima. Melanie II [7] subtracts the minimum intensity from all pixel values and then fits a third degree polynomial to the background image (with spots removed). Another technique is derived from 3-D mathematical morphology, where the operations of opening and closing a grayscale image by a structuring element is represented by sliding the structuring element respectively under and over the topographical image (intensity is regarded as height) [9]. Moreover, using a horizontal and/or vertical cylindrical structuring element, horizontal or vertical streaks may also be removed [10].

Previous work in 2DGE image segmentation includes several single-phase direct segmentation methods that will be reviewed here. They include methods using stepwise thresholding [14], second derivatives [15], the Watershed transform [7], and statistical spot modeling methods [13]. The stepwise thresholding approach is extremely sensitive to noise and artifacts, where additional criteria must hold in order to accept or reject the final connected areas. The second derivatives approach gives acceptable results only when proper noise suppression has been applied. Furthermore, it places the borders at the inside of the spots since the zero crossings of the second derivative are associated with the steepest part of the spot rather than its beginning. The Watershed transform based method has the major disadvantage of over-segmentation. Although this can be addressed using marker controlled watersheds, the selection of a good set of markers is not a trivial task. Finally, the approaches using statistical spot modeling are difficult to apply without prior knowledge of spot shapes and sizes and it is known that they perform poorly in areas with overlapping spots especially if these foreground areas are not accurately estimated (usually this estimation is performed by mathematical morphology). So, it is obvious then that the current protein spot detection approaches suffer from various disadvantages, such us sensitivity to noise and artefacts, spot border distortions, over-segmentation, poor performance in areas with overlapping spots [4] etc. Furthermore, they require careful post-processing and usually a lot of manual effort, to finally produce reliable detection results.

1.2 Dissertation Contributions

The goals of this dissertation have been the development of novel methods for 2DGE image analysis and especially for spot denoising, detection and quantification in order to improve the accuracy and efficiency of existing methods. Image analysis is still a bottleneck in expression proteomics workflows. Nowadays, there are several commercial software packages available such as PDQuest, ImageMaster, Progenesis, etc, that promise to be accurate and efficient but this is far from being a reality [16]. Motivated by the aforementioned limitations the main contributions are:

✓ *2DGE image denoising*

Since 2DGE gel images are inherently noisy due to dust and the imperfect image acquisition process, the first objective of this dissertation is the development of an effective denoising method, i.e. increasing the SNR without inserting significant distortions to the image. In this dissertation, a multiresolution image transform was employed, namely the Contourlet Transform (CT), which proved to be very effective for denoising 2DGE images and fit well the specific properties of gel images. The CT can approximate more accurately images with smooth contours and anisotropic characteristics. 2DGE images are anisotropic due to the large variety in shapes and orientation of the spots they contain. The developed method is fully automated and it is shown to outperform every previously reported method [17,18]. We must note that the CT has not been used before for this type of images nor it has been coupled with the coefficient thresholding techniques that we have used.

- ✓ *Novel Active Contours based method for extracting foreground Regions of Interest (ROIs)*

A novel methodology [19] has been developed for delineating 2DGE image areas which, with high probability, include protein spots, based on Active Contours without Edges (ACWE). Moreover, a technique based on Contourlet Transform has been developed that leads to the automatic determination of the initial curve. This initialization method reduces the convergence time of the algorithm and improves its efficiency. Due to fact that the Contourlet Transform has properties that match particular characteristics of 2DGE images, a method based on it has been developed in order to enhance gel images and especially the faint spots. The method has been evaluated using a large pool of synthetic and real gel images. It has been shown that the extracted ROIs include the large majority of the true spots and are tight, i.e. they do not include large background areas. The evaluation has been performed using the popular commercial software package PDQuest and also relatively to the provided ground truth. Furthermore, our method does not require re-calibration of parameters every time a new image is processed and it can thus be fully automated.

- ✓ *Novel hierarchical approach for protein spot detection & quantification using machine learning methods*

This approach [20,21], unlike the traditional spot detection workflow where a gel image is directly segmented into spot regions following the spot modelling phase, it is applied on each ROI resulting from the previous image analysis step. First, it removes the local background pixels and streaks using 1-dimensional Gaussian mixture models applied on the intensity histograms of the extracted ROIs. Unlike mathematical morphology filtering it “kills” streak pixels without affecting the true spot pixels. A key idea of the developed method is that the informative image pixels are treated as sample data generators where machine learning methods are applied to the so generated data samples. A core technique used repeatedly in the developed methodology is Gaussian Mixture Modelling (GMM) [22] which is applied in an unsupervised manner [23]. Through an extensive evaluation, we have demonstrated that the developed methodology achieves trustworthy detection results and introduces much less spot artifacts. In addition, a comparison with the popular commercial software package PDQuest was conducted and shown that the developed methodology is more precise and more specific than PDQuest, while both methods achieve high sensitivity. Furthermore, it has been shown that it leads to more accurate spot quantification than PDQuest. Finally, the developed methodology can be fully automated and thus it is labor and error free from the user’s perspective, which is very important for high throughput proteomics projects.

2 Image Analysis

2.1 Image Denoising

The denoising methods commonly used so far, have the tendency to deform the protein spots on the gel to the extent that they create extraneous spots i.e. artifacts. This is a serious problem since insufficient or improper denoising affects the whole image processing pipeline from its early stages. So, it impacts negatively all the subsequent processes, such as spot detection, spot quantification, as well as spot matching across gels. In order to surpass those problems, a novel method for denoising 2DGE images has been developed, based on the Contourlet Transform [12]. The Contourlet Tranform (CT) is a multiresolution, flexible, directional image decomposition method based on contour segments. The main difference between the CT and the WT is that the CT allows for a different number of directions at each scale (can be any power of 2). So, CT can represent more efficiently smooth spot contours in a 2DGE image.

The denoising by multiresolution transforms involves the aforementioned analysis of signal followed by a coefficient thresholding method (also called *shrinkage*). For the developed CT-based

2DGE denoising methodology, two of the best performing shrinkage methods reported in the WT literature were adopted, namely the *BayesThres* [24] and *Bivariate* [25] methods.

As it is demonstrated the Contourlet Transform has properties that match well the characteristics of 2DGE images, and after a thorough evaluation with both synthetic and real gel images in terms of the achieved Signal to Noise Ratio (SNR), the distortions introduced and mainly via the benefits it offers to subsequent image analysis steps:

1) Protein spot detection - where by using the developed denoising methodology we avoid introducing a large number of artifacts and detect more faint spots.
2) Protein spot quantification - the estimated spot quantities are more close to the known ground truth and with less variance than when using wavelet-based denoising.

In conclusion, the developed denoising methodology is more effective than the currently used spatial filtering methods implemented by commercial software packages and also more effective than the more recently introduced Wavelet-based denoising methods.

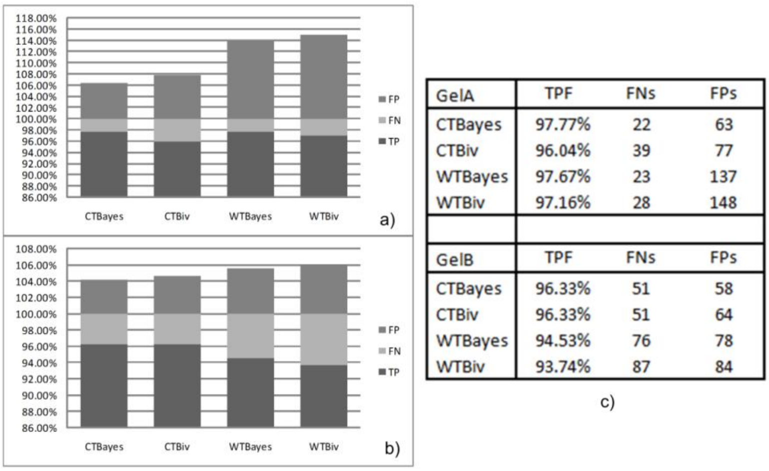
Next we present the results of the developed method and compare them with the currently state-of-art method (wavelet denoising). We used PDQuest (version 8.0.1) to evaluate the different denoising approaches in terms of spot detection achieved after denoising using real images. We evaluated each method in terms of the TPs, FPs (artifacts) and false negatives (FNs) or missed spots. The results are summarized in Figure 1. We notice that regardless of the image used, the CT based denoising approaches result to a considerably smaller percentage of introduced extraneous spots (FPs), ranging from 4% to 8%, compared to the WT based denoising methods where the corresponding range was from almost 6% to 15%. The missed spots (FN) were also less when using the CT in all cases except for the CT-Bivariate and GelA case where they approached 4%. Overall, CT-Bayes denoising outperformed all other methods in terms of TPs, FNs and FPs and for both real images used (see Figure 1(c)).

2.2 Active Contours based method for extracting foreground Regions of Interest (ROIs)

Segmentation of 2DGE images requires partitioning them into areas of foreground (include protein spots) and background (no protein spots). We developed a new method based on Active Contours [26] that separates effectively those two areas in a way that: (i) reduces the number of missed faint spots, (ii) finds correct and tight borders for areas with spots, (iii) avoids over-segmentation.

Active contours (ACs) are a very powerful tool for image segmentation and object tracking. The key idea is the evolution of a curve, or curves, also called "snakes", subject to constraints from the input image. Due to the properties of the specific application evolving curves should allow automatic topological changes of the curve. An AC approach that holds this property was introduced in [27] where the curve is modelled as a specific level set function of time in a higher dimensional surface. For more details on the developed methodology reader is referenced to [19].

Next we present some results of the developed methodology while for more details the reader is referenced to [19]. The results have also been compared to the ones obtained with PDQuest. The denotation followed is: spots that AC missed but were found by PDQuest (PDQ/nAC) partitioned into two subsets; existing spots that our AC base method missed (false negatives, FN) and extraneous spots (true negatives, TN = PDQuest artefacts) that AC correctly ignored. For those spots that AC detect but PDQuest missed (denoted as nPDQ/AC) we also consider two subsets. The ones that AC correctly found because they do exist (true positives, TP_2), and those that AC found but do not exist (false positive, FP = AC artefacts). The results obtained are summarized in Tables 1 and 2. Knowing that PDQuest is performs very well in segmentation we can conclude that our approach correctly reports the foreground regions in a wide collection of different of 2DGE images. We also see that we succeed in avoiding some PDQuest detected artefacts (TN) but also fail to detect very few spots that should have been detected (FN). On the other hand, we detect areas that contain spots missed by PDQuest (TP_2), but also introduce some artefacts (FP). In Table 1 we can see that the ratio of spots missed by our approach compared to PDQuest is pretty small (<3%, except for the Rj1 image) which also indicates that the proposed method results are highly reliable. In addition, as we can see from Table 2, that the proposed approach achieves sensitivity over ~91% for all images and a confidence above ~96%. These results indicate that ACs can be very effective in confining protein spots into tightly bounded spot areas. Accurate and correct spot areas segmentation is a prerequisite for spot detection and quantification. We have shown that the proposed AC based segmentation achieves comparable results with a mature tool for 2DGE image analysis (PDQuest) but with much less user intervention.



GelA	TPF	FNs	FPs
CTBayes	97.77%	22	63
CTBiv	96.04%	39	77
WTBayes	97.67%	23	137
WTBiv	97.16%	28	148
GelB	TPF	FNs	FPs
CTBayes	96.33%	51	58
CTBiv	96.33%	51	64
WTBayes	94.53%	76	78
WTBiv	93.74%	87	84

Figure 1 Spot detection results using the real images from [27] a) GelA and b) GelB respectively, c) the corresponding True Positive Fraction (TPF), False Positives (FPs) and False Negatives (FNs).

Image	PDQ	PDQ/AC	% PDQ/AC	PDQ/nAC	% PDQ/nAC	nPDQ/AC	%nPDQ/AC
1a	1112	1090	98,02%	22	1,98%	13	1,19%
2a	1315	1283	97,57%	32	2,43%	7	0,55%
MP1	262	256	97,71%	6	2,29%	13	5,08%
MP2	265	242	91,32%	23	8,68%	11	4,55%
MP3	227	223	98,24%	4	1,76%	24	10,76%
Rj1	146	123	84,25%	23	15,75%	4	3,25%
RGA	948	919	96,94%	29	3,06%	9	0,98%
RGB	1040	1018	97,88%	19	1,83%	40	3,93%

Table 1 Evaluation results. PDQuest spots in AC extracted foreground regions were all real spots (PDQ/AC = TP_1). We also report spots detected by PDQuest and not included in our foreground areas (PDQ/nAC = (TN+FN)) and spots in our foreground areas not detected by PDQuest (nPDQ/AC = (TP_2+FP)).

Image	PDQ	AC/PDQ (TP_1)	PDQ/nAC		nPDQ/AC		S	C
			FN	TN	FP	TP_2		
1a	1112	1090	5	17	5	8	99,55%	99,55%
2a	1315	1283	9	23	5	2	99,30%	99,61%
MP1	262	256	2	4	10	3	99,23%	96,28%
MP2	265	242	14	9	4	7	94,68%	98,42%
MP3	227	223	1	3	6	18	99,59%	97,57%
Rj1	146	123	11	12	1	3	91,97%	99,21%
RGA	948	919	20	9	6	3	97,88%	99,35%
RGB	1040	1018	12	7	13	27	98,86%	98,77%

Table 2 Evaluation results. Sensitivity is above 91% and Confidence above 96% for all images.

2.3 Protein Spot Detection & Quantification

Previously developed methods for protein spot detection suffer from various disadvantages, such us sensitivity to noise and artifacts, spot border distortions, over-segmentation and poor performance in areas with overlapping spots [4]. Furthermore, they require careful post-processing and a lot of manual effort, to finally produce reliable detection results [16]. To address these limitations, a novel approach for 2DGE automatic spot detection and quantification has been developed. Here, the informative ROI pixels are treated as sample data generators and Gaussian Mixture Modeling (GMM) is applied in an unsupervised manner [29,30], i.e. no pre-training is required and the whole process can be fully automated.

2.3.1. Local Background and Streaks removal

Although the ROI extraction step results in areas that include the vast majority of the protein spots present in a gel image, they may also include some local image background pixels and/or streak segments. In order to eliminate those pixels that may confuse spot modeling downstream from further processing, we classify the object pixels into 3 possible classes: class-1 represents the "strong" and/or saturated spot pixels, class-2 the "faint" spot pixels and tails of "strong" spots, and finally class-3 correspond to the pixels of the local background and/or streaks. The classification is performed on the histogram of pixel intensities using 1-dimmensional Gaussian mixture modeling and a modified Expectation-Maximization (EM) algorithm [30]. The Minimum Message Length (MML) criterion (proposed in [30] for model selection) is applied to determine the optimal number of components in the mixture model that best fits the histogram data.

In summary, the use of the modified EM fits 3 pixel classes to the histogram while the MML criterion tests this fit and discards any class that is not essential for the histogram's interpretation. So, if the histogram is well explained by 3 classes we can then threshold out the class-1 pixels that correspond to the background and/or streaks (light gray in Figure 3 (3)), else (if we end up with less than 3 classes) we keep the object intact. As one can notice, the developed method is performing local background and streak removal task only in areas and to the extent that it is really needed.

2.3.2 Initial Spot centers Estimation

At the next step we estimate the number of protein spots existing inside each ROI. To do so, a 5x5 (pixels) spatial filter is employed that finds the local minima (zero intensity corresponds to black pixels, maximum intensity to white) in the image (see green asterisks in Figure 3 (5)). Due to the pixel intensity saturation effect, it is possible that the filter identifies several closely located local minima. These are actually replicates of the same candidate spot centre and need to be grouped appropriately so as we do not end up with a very large and misleading number of candidate spot centers per object. This is accomplished by applying agglomerative Hierarchical Clustering (HC) [30] of the minima points using the Manhattan distance, the single linkage method for merging formed clusters. The extracted candidate spot centers are depicted by red circles in Figure 3(5).

2.3.3 2D Spot Modeling

The next and most distinguishing characteristic of our pipeline is the idea of using the pixel intensities as data generators. The total number of data points *N* generated by random sampling for each ROI is kept proportional to the number of estimated candidate spot centres, and not to the area of the ROI. The *N* points to be drawn are distributed among the pixels of the object according to their relative intensities (a "stronger" pixel "throws" more points in its neighbourhood). This is in accordance to the view that pixel intensity ideally represents the quantity of protein molecules concentrated at that particular gel location. Specifically, each pixel *i* with intensity I_i acts as the centre $\mu=(x_i,y_i)$ of a 2D Gaussian component $N(\mu,\Sigma)$ in a Gaussian Mixture Model [28] having as many components as the number of pixels and a predetermined fixed covariance matrix Σ.

As we can notice from Figure 3 (6) (light grey data points) the generated set of data points may include points that are far from all candidate spot centres. These outlier points (due to remaining background pixels) can be identified since they have low likelihood for all components of the mixture (no component “feels strongly” about them) and are removed at this stage since they may adversely affect the subsequent step of spot detection and quantification.

The last step in the pipeline is the application of Gaussian Mixtures Models (GMM) [28] in 2-dimensions (see figure 3(7) & 3(8).

2.3.3 Results & Discussion

2.3.3.1 Spot detection & Quantification evaluation

In this Section the results of the developed methodology are presented. Figure 4 presents a summary of the results. It can be noticed that the developed method achieves a high TPF (over 91% for all images) and a high PPV (over 79% for all images), meaning that it is both sensitive and precise. This conclusion is also supported by the fact that the images exhibit very different characteristics.

This is because these images exhibit a larger dynamic range difference between spots and background. So, Active Contours based segmentation performs better at complex areas leaving out of the ROIs only a small portion of faint spots, unlike PDQuest which in order to achieve high detection efficiency needs a larger sensitivity parameter value which leads to a very large number of extraneous artefacts (greater than 300 artefacts).

Quantification of protein spots is also a very crucial step in 2-DE image analysis. Its accuracy and reliability greatly affects the subsequent differential spot expression analysis. Figure 5(a) presents the results obtained for the noise free synthetic image for which the ground truth (spot volumes) is known. This is a scatter plot of estimated vs. ideal spot volumes, as produced by PDQuest (data points denoted by circles) and the developed method (data points denoted with an x). The ideal regression line would be the diagonal line $y=\alpha x$ with $\alpha=1$. As it can be seen in Figure 5(a), the regression lines for each method (dotted line for PDQuest and solid line for developed method) deviate from the ideal. In the case of PDQuest, where $\alpha=1.906$, it is clear that there is an overestimation of the true spot volumes, which according to the table in Figure 5(b) (first column) has the mean value of 23.6 units. On the other hand, the developed method has $\alpha=0.8655$ i.e. it underestimates the true volume with a mean error of -4.6 units. Someone could argue that since the bias introduced by each method affects all the spots in a gel image it is of no great concern to differential expression analysis. However, regardless of the type of bias inserted by each method, the most important result is the error’s standard deviation (see Figure 5 (a) and (b)). As we can see, the developed methodology exhibits a low standard deviation of 0.42 while the STD of PDQuest estimated spot volumes for the same image is more than ten times higher (6.38). The corresponding Root Mean Square Error (RMSE) is 4.77 and 30.03 respectively. The much lower STD and RMSE values of quantification error suggest that when using the developed methodology the variance of the produced quantity estimates across spots with similar characteristic is much smaller. So, the developed approach is highly consistent, in the sense that a gradual decrease in spot maximal intensity (which also reflects in the spot’s area in this image) leads to a corresponding gradual decrease in volume estimates (see Figure 5(a)). This characteristic is what is the most important in practice for a reliable differential analysis based on spot volumes, irrespectively of any bias introduced by the method. Finally, at Figure 5(e) someone can see how noise addition and subsequent denoising affects spot detection performance for each method. It is clear that the developed method exhibits robustness to noise in detecting faint spots and avoiding introducing artefacts. As someone can notice, the developed methodology exhibits a high degree of robustness in detecting faint spots. Furthermore, it avoids introducing artifacts. It should also be mentioned that the developed method does not need any user intervention while PDQuest requires careful selection of its detection parameters every time the image is changing.

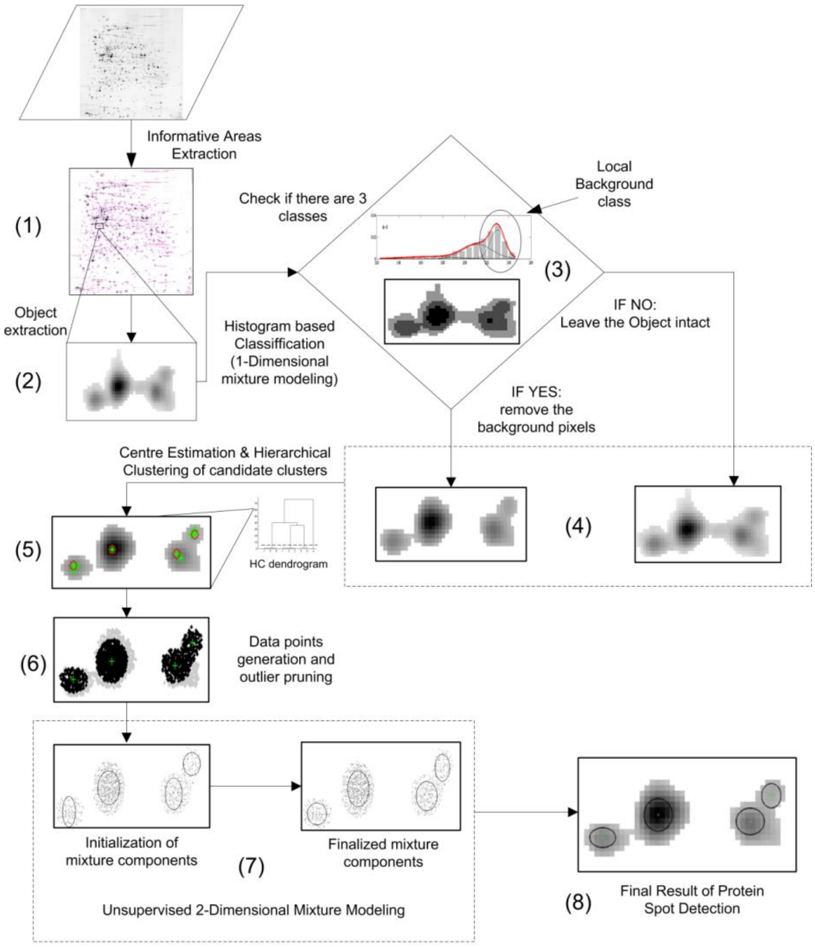
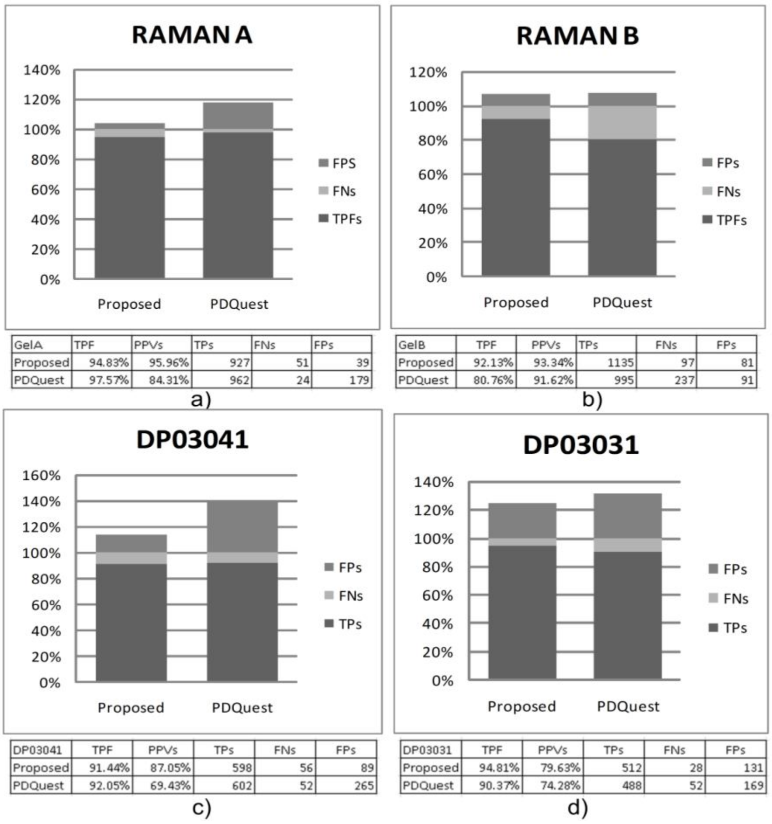

Informative Areas Extraction
(1)
Object extraction
(2)
Histogram based Classiffication (1-Dimensional mixture modeling)
Check if there are 3 classes
Local Background class
(3)
IF NO: Leave the Object intact
IF YES: remove the background pixels
Centre Estimation & Hierarchical Clustering of candidate clusters
(4)
(5)
HC dendrogram
(6)
Data points generation and outlier pruning
Initialization of mixture components
(7)
Finalized mixture components
Unsupervised 2-Dimensional Mixture Modeling
(8)
Final Result of Protein Spot Detection

Figure 3: Overview of the developed spot detection method. Panels: **(1)** Extraction of areas containing protein spots using Active Contour based first level Segmentation, **(2)** Extracted image object, **(3)** Histogram based classification using 1-Dimensional mixture modelling (upper image shows the histogram of the object intensities, red line indicates the joint mixture density consisting of the individual densities represented with solid lines (1-D Gaussians); bottom image presents the resulting classified regions where the light gray area indicate the third class (local background) to be eliminated), **(4)** Resulting object after the histogram classification, **(5)** Centre estimation (green asterisks) and merging (red circles) of the neighbouring replicate centres with hierarchical clustering (dotted line at the dendrogram on the right), **(6)** Data generation and manipulation, **(7)** Unsupervised 2-Dimensional mixture modelling and final estimate, **(8)** Final detection result on the image object.

3 Conclusions

The goals of this dissertation have been the development of novel methods for 2DGE image analysis and especially for spot denoising, detection and quantification in order to improve the accuracy and efficiency of existing methods. Image analysis is still a bottleneck in expression proteomics workflows. Nowadays, there are several commercial software packages available such as PDQuest, ImageMaster, Progenesis etc, that promise to be accurate and efficient but this is far from being a reality [16]. In order to surpass current limitations we developed an end-to-end image processing pipeline which addresses effectively the current bottlenecks of the available 2DGE image analysis methods. We proved its robustness and efficiency using real and synthetic datasets while also evaluated its performance by means of what matters most for the proteomics scientists more than for engineering scientists. Finally, the developed methodology can be fully automated and free the user from the time consuming tasks of parameter fine tuning and spot editing.



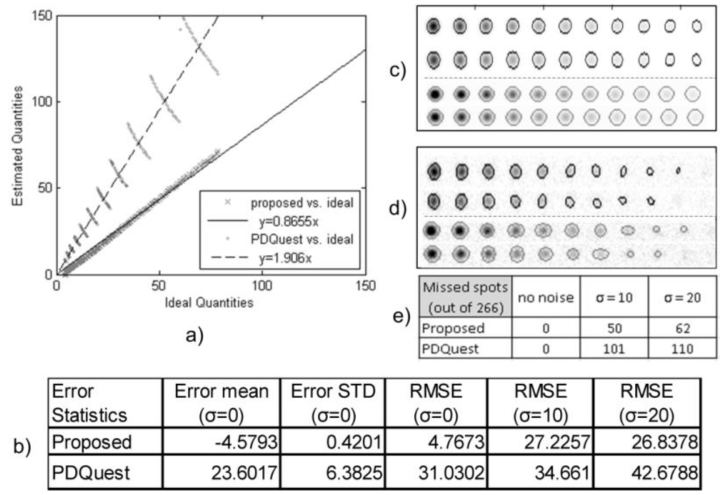
GelA	TPF	PPVs	TPs	FNs	FPs
Proposed	94.83%	95.96%	927	51	39
PDQuest	97.57%	84.31%	962	24	179

GelB	TPF	PPVs	TPs	FNs	FPs
Proposed	92.13%	93.34%	1135	97	81
PDQuest	80.76%	91.62%	995	237	91

DP03041	TPF	PPVs	TPs	FNs	FPs
Proposed	91.44%	87.05%	598	56	89
PDQuest	92.05%	69.43%	602	52	265

DP03031	TPF	PPVs	TPs	FNs	FPs
Proposed	94.81%	79.63%	512	28	131
PDQuest	90.37%	74.28%	488	52	169

Figure 4 a) Comparison of results for the image *RamanA* using the developed method and the PDQuest software package in terms of True Positives (TPs), False Negatives (FNs) and False Positives (FPs); **b)** for image *RamanB*; **c)** for image *DP03041*; **d)** for image *DP03031*.



Missed spots (out of 266)	no noise	σ = 10	σ = 20
Proposed	0	50	62
PDQuest	0	101	110

Error Statistics	Error mean (σ=0)	Error STD (σ=0)	RMSE (σ=0)	RMSE (σ=10)	RMSE (σ=20)
Proposed	-4.5793	0.4201	4.7673	27.2257	26.8378
PDQuest	23.6017	6.3825	31.0302	34.661	42.6788

Figure 5: Comparative evaluation of the developed method to PDQuest for spot quantification performance. **a)** Estimated vs. ideal spot quantities using the synthetic image *Quant*. **b)** The table provides the quantity error means and standard deviations in the case of noise free image along with the corresponding root mean square error (RMSE) for all noise cases. **c)** Image patches with detected spots from the noise-free image (the top two spot rows correspond to results of the developed method, while the bottom two spot rows to PDQuest results). **d)** Image patches with detected spots from the noise corrupted image with *σ=10* (top two spot rows correspond to results of our method while the bottom two spot rows to PDQuest results). **e)** Spot detection performance (missed spots) of each method after denoising, for high and very high noise levels.

4 References

[1] Anderson NL, Anderson NG, "Proteome and proteomics: new technologies, new concepts, and new words", Electrophoresis 19 (11): 1853–61, 1998.

[2] Blackstock WP, Weir MP, "Proteomics: quantitative and physical mapping of cellular proteins", Trends Biotechnology, 17 (3): 121–7, 1999.

[3] Marc R. Wilkins, Christian Pasquali, Ron D. Appel, Keli Ou, Olivier Golaz, et al., "From Proteins to Proteomes: Large Scale Protein Identification by Two-Dimensional Electrophoresis and Amino Acid Analysis". Nature Biotechnology 14 (1): 61–65, 1996.

[4] Dowsey, A. W., High-throughput image analysis for proteomics. PhD Thesis 2005, Department of Computing, Imperial College London.

[5] Pietrogrande, M. C., Marchetti, N., Dondi, F., Righetti, P. G., "Spot overlapping in two-dimensional polyacrylamide gel electrophoresis separations: a statistical study of complex protein maps", Electrophoresis, 2002, 23, 283-291.

[6] Kaczmarek, K., Walczak, B., de Jong, S., Vandeginste, B. G. M., Preprocessing of two-dimensional gel electrophoresis images. Proteomics 2004, 4, 2377-2389.

[7] Appel, R. D., Vargas, J. R., Palagi, P. M., Walther, D., et al., "Melanie II – a third-generation software package for analysis of two-dimensional electrophoresis images: II. Algorithms", Electrophoresis, 1997, 18, 2735-2748.

[8] Tyson, J. J., Haralick, R. H., "Computer analysis of two-dimensional gels by a general image processing system", Electrophoresis, 1986, 7, 107-113.

[9] Sternberg, S. R., "Gray scale morphology", Comp. Vis. Graph. Image Processing, 1986, 333-355.

[10] Skolnick, N. M., "Application of morphological transformations to the analysis of two-dimensional electrophoresis gels of biological materials", Comput. Vis. Graph. Image Process., 1986, 35, 306-322.

[11] M. Rogers, J. Graham, and R.P. Tonge, ”Using statistical image models for objective evaluation of spot detection in two-dimensional gels”, Proteomics, 2003, vol. 3, pp. 879- 886.

[12] M.N. Do and M. Vetterli, ”The contourlet transform: An efficient directional multiresolution image representation”, IEEE Trans. on Image Proces., 2005, vol. 14, pp. 2091-2106.

[13] Conradsen, K., Pedersen, J., "Analysis of two-dimensional electrophoretic gels", Biometrics, 1992, 48, 1273-1287.

[14] P. Cutler, G. Heald, I. R. White, J. Ruan, “A novel approach to spot detection for two-dimensional gel electrophoresis images using pixel value collection”, Proteomics, vol. 3, pp. 392-401, 2003.

[15] Lemkin, P. F., Lipkin, L. E., "GELLAB: a computer system for two dimensional gel electrophoresis analysis. III. Multiple two-dimensional gel analysis", Comput. Biomed. Res., 1981, 14, 407-446.

[16] Clark, B. N., and Gutstein, H. B., The myth of automated, high-throughput 2DGE image analysis, Proteomics 2008, 8, 1197-1203.

[17] Tsakanikas, P., Manolakos, E. S., Improving 2-DE gel image denoising using Contourlets, Proteomics 2009, 9, 3877–3888.

[18] Tsakanikas, P., Manolakos, E. S., Effective Denoising of 2D Gel Proteomics Images Using Contourlets IEEE Proc. ICIP 2007, San Antonio, Texas, USA, September 16-19, 2007, pp. VI: 269-272.

[19] Tsakanikas, P., Manolakos, E. S., Active Contour Based Segmentation of 2DGE Proteomics Images, 16th European Signal Processing Conference (EUSIPCO-2008), Lausanne, Switzerland, August 25-29, pp. 83-87, 2008.

[20] Tsakanikas, P., Manolakos, E. S., Protein Spot Detection and Quantification in 2-DE gel images using Machine Learning methods, Proteomics 2011, accepted, to appear.

[21] Tsakanikas, P., Manolakos, E. S., “A fully automated 2-DE gel image analysis pipeline for high throughput proteomics”, International Conference on Acoustics, Speech and Signal Processing (ICASSP 2011), May 22-27, 2011, accepted, to appear.

[22] McLachlan, G.J. and Peel, D. Finite Mixture Models, Wiley (2000).

[23] Figueiredo M., and Jain A.K., Unsupervised learning of finite mixture models, IEEE Transactions on Pattern Analysis and Machine Intelligence - PAMI, vol. 24, no. 3, pp. 381-396, March 2002.

[24] Abramovitch, F. F., Sapatinas, T., Silverman, B., Wavelet thresholding via a Bayesian approach, J. R. Stat. 1998, 60, 725-749.

[25] Sendur, L., Selesnick, I. W., Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency. IEEE Trans. on Signal Processing 2002, 50, 2744-2756.

[26] T. F. Chan, L. A. Vese, “Active Contours without Edges”, IEEE Trans. on Image Processing, vol. 10, pp. 266-277, Feb. 2001.

[27] S. Osher, J. A. Sethian, “Front Propagating with curvature dependent speed: Algorithms based on Hamilton-Jacobi Formulation”, J. Comput. Phys., vol. 79, pp. 12-49, 1998.

[28] Raman, B., Cheung, A., Marten, M. R., Quantitative comparison and evaluation of two commercially available, two-dimensional electrophoresis image analysis software packages, Z3 and Melanie. Electrophoresis 2002, 23, 2194-2202.

[29] McLachlan, G.J., Peel, D. Finite Mixture Models,Wiley 2000.

[30] Figueiredo M., and Jain A.K., Unsupervised learning of finite mixture models, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 24, no. 3, pp. 381-396, March 2002.

[31] Theodoridis, S. and Koutroumbas, K., Pattern Recognition (Third Edition), Elsevier, 2006