CABIOS

Vol.9 no.2. 1993
Pages 133 – 139

A robust, high-sensitivity algorithm for automated detection of proteins in two-dimensional electrophoresis gels

Jerry E.Solomon and Michael G.Harrington

Abstract

The automated interpretation of two-dimensional gel electrophoresis images used in protein separation and analysis presents a formidable problem in the detection and characterization of ill-defined spatial objects. We describe in this paper a hierarchical algorithm that provides a robust, high-sensitivity solution to this problem, which can be easily adapted to a variety of experimental situations. The software implementation of this algorithm functions as part of a complete package designed for general protein gel analysis applications.

Introduction

The primary motivation for the work described in this paper has been development of automated methods for robust, high-sensitivity analysis of two-dimensional gel electrophoresis (2DGE) images of protein samples. The 2DGE protein separation method provides a powerful technique for protein research in molecular biology and medical diagnostics by allowing separation of thousands of proteins and polypeptides according to charge and molecular weight in an image format (O'Farrell, 1975; Garrels, 1989; Skolnik *et al.*, 1982). Two examples of such gel imagery are shown in Figure 1, where the vertical dimension is proportional to molecular weight and the horizontal dimension is proportional to molecular charge. In this image the pixel intensity value is inversely proportional to gel optical density, i.e. black indicates high protein concentration. Since ten to hundreds of such gel images must be analyzed in a typical experiment or clinical study, it becomes important to develop computer-automated methods for such analyses.

Protein 2DGE images exhibit quite distinctive characteristics according to sample type; Figure 1(a) represents a typical cellular protein pattern, while Figure 1(b) illustrates the more heterogeneous appearance exhibited by body fluid proteins. Comparisons of similar samples are of primary interest in biology and medicine, but there are two noteworthy aspects of this technology: if the same source of protein is applied to two consecutive gels, the resultant image has both general technical consistency and subtle degrees of technical variation. The predominant consistency of such an image has led to identification of unique diagnostic individual protein changes, even such minor changes as those in Figure 1(b) from cerebrospinal fluid (CSF) of a patient with a specific transmissible dementing disease; all such patients have the presence of the proteins marked by arrows in the figure, whereas all patients with other causes of dementia, such as Alzheimer's disease, do not have these proteins present in their CSF (Harrington *et al.*, 1986). These small yet unique changes illustrate the > 100-fold increase in ability to characterize biological/diagnostic data compared to previous methods, but their accurate identification presents an analysis challenge of some magnitude. In contrast with this level of overall reproducibility, the more subtle technical variations result from inconsistencies in the electrophoretic separation and detection processes, and lead to occasional distortion of relative positions, shape and intensities of spots. Efforts to improve technical reproducibility are continuing, but the problem will remain a challenge to image analysis methodologies because of the interaction of the innumerable components during the electrophoretic process.

A fundamental image analysis problem to be addressed is that of identifying and characterizing those intensity distributions in the image that represent actual protein distributions in the original gel. As can be seen from the images in Figure 1, the protein distributions may be described as intensity blobs having highly variable shape characteristics, diffuse edges and a wide range of peak intensities. In addition, these images typically have regions in which whole families of proteins are tightly grouped, resulting in overlapping intensity distributions that image analysis operations must attempt to resolve. Since trace proteins are quite often the ones of most interest, both in biological experiments and in clinical studies, sensitivity and false alarm performance of any detection method are serious issues.

We should emphasize that the current work has concentrated on developing a robust method that allows reliable spot detection of 'faint spots' on lightly loaded gels. Routine use of this algorithm in our laboratory for the past two years indicates that it works equally well under relatively 'heavy' gel loading conditions. It should also be pointed out that we have not attempted an exhaustive comparison of our method with many of the other spot finding techniques currently in use within the general community since most, though not all, rely in one way or another on prior information about spot shapes such as two-dimensional Gaussian, etc.

Division of Biology and The Beckman Institute, California Institute of Technology, Pasadena, CA 91125 USA

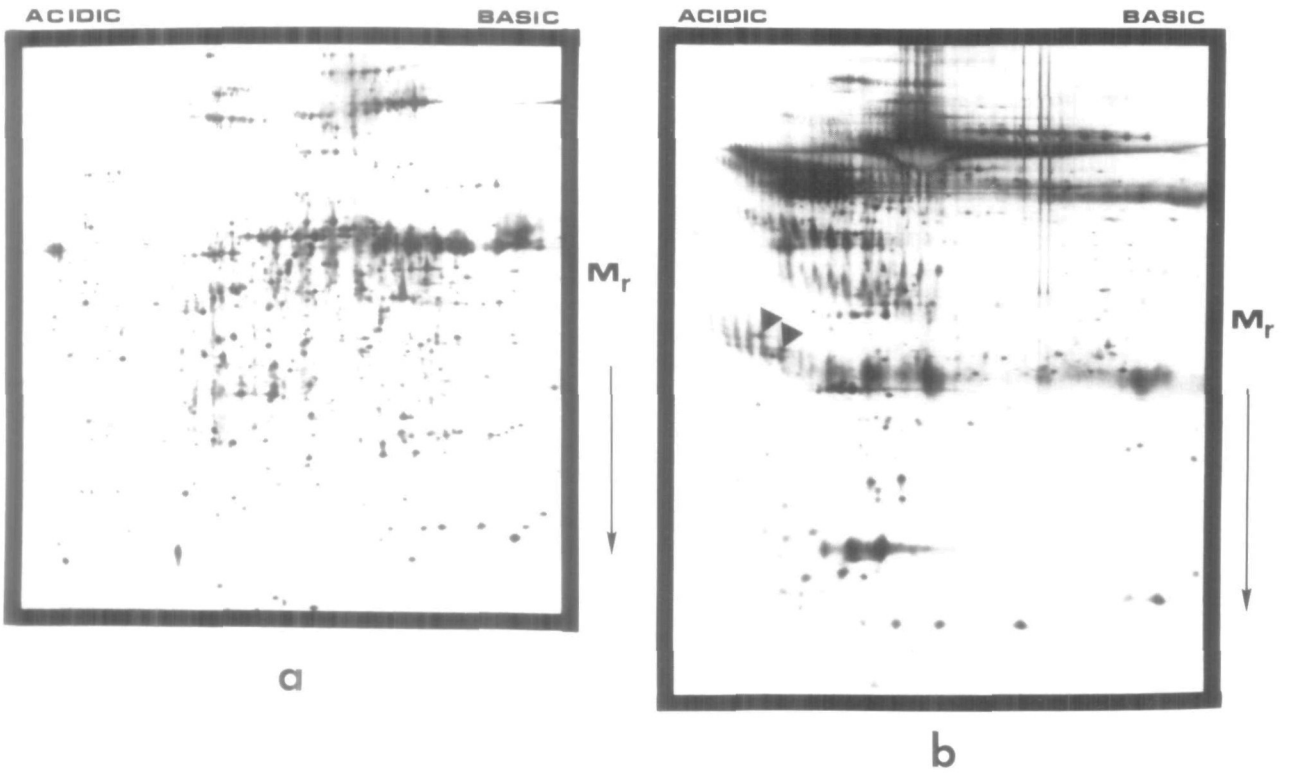

Fig. 1. Two examples of digitized silver-stained 2DGE images: (a) typical appearance of proteins derived from cellular material, (b) proteins derived from CSF.

System and methods

The protein spot detection and characterization algorithm was developed in a Sun Workstation Unix environment (SunOS version 4.1) as one part of a complete 2DGE analysis software package. The algorithm is implemented in standard C language source code and has been compiled and tested on Silicon Graphics and Alliant multiprocessor machines, as well as on a variety of Sun-3, Sun-4 and Sparcstation machines. As is the case in all large-scale image analysis applications, ample amounts of local memory (RAM) are required for efficient operation. For large format image data a minimum of 16 Mbytes is required; however, 24–32 Mbytes is recommended for efficient operation. Both silver-stained gels and gel autoradiograms are digitized with a laser scanning densitometer with spatial sampling rate of 80 μm, and 12 bits of quantization in optical density.

Algorithm description

A major problem that continues to plague developers of automated digital image analysis systems is the robust detection and characterization of objects delineated by diffuse boundaries; generally referred to as fuzzy objects, or blobs. If the objects of interest can be well characterized by regular convex boundaries enclosing a single intensity maximum, then peak-finding/boundary estimation methods, such as those developed by O'Gorman and Sanderson (1984, 1986), may usually be applied. In the case of 2DGE images, a common approach has been to model the spots as two-dimensional Gaussian density distributions and use fitting or template matching methods for detection as discussed by Garrels (1989). However, there are many cases in which protein spots are not well modeled as Gaussian distributions, and this approach does not therefore provide a general solution. In more complex cases one must in general resort to an attack based on edge detection methods, which immediately raises the issue of robust detection of diffuse (or fuzzy) edges. Most edge detectors are designed using a optimality criterion based on step-edge detection, rather than diffuse edge properties (see Canny, 1986, for a review). Thus, in the case of fuzzy objects, one generally seeks a preprocessing operator that transforms diffuse edges into edges that more closely resemble step-edges in order to utilize the optimal properties of known edge detection algorithms. Once a robust method for diffuse edge detection is found one must then have a reliable method for selecting those edge-bounded objects of interest, i.e. in this case density distributions that truly represent proteins. We have developed a two-stage, hierarchical approach to this problem which achieves very high sensitivity detection in the first stage, while applying an efficient set pruning heuristics at the second stage in order to reject false positive detection events.

Edge detection stage

The first step in detecting protein spots is to delineate the diffuse edges which constitute potential spot boundaries. The method described here is based on use of the so-called Laplacian-of-Gaussian (LOG) $\nabla^2 G$ operator, the properties of which are reviewed in Marr and Hildreth (1980), Torre and Poggio (1986) and van Vliet *et al.* (1989). When considered as a convolution operator, the LOG kernel has the form:

$$\nabla^2 G_\sigma(x,y) = \frac{1}{2\pi\sigma^4}\left[2 - \left(\frac{(x^2+y^2)}{\sigma^2}\right)\right]\exp\left[-\frac{(x^2+y^2)}{2\sigma^2}\right] \quad (1)$$

where σ determines the spatial scale of the operator through the relation $w = 2\sqrt{2}\sigma$; w being the extent of the positive central region of the operator. As shown by Huertas and Medioni (1986), this operator is separable, and computations may be carried out as two successive one-dimensional convolutions using the prescription:

$$\nabla^2 G_\sigma(x,y) = h_{12}(x,y) + h_{21}(x,y) \quad (2)$$

with

$$h_{12}(x,y) = h_1(x)\cdot h_2(y)$$

$$h_{21}(x,y) = h_2(x)\cdot h_1(y)$$

$$h_1(\zeta) = \frac{1}{\sqrt{2\pi\sigma^2}}\left(1-\frac{\zeta^2}{\sigma^2}\right)\exp\left[-\frac{\zeta^2}{2\sigma^2}\right]$$

$$h_2(\zeta) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\zeta^2}{2\sigma^2}\right]$$

Edge detection is traditionally carried out by using a heuristic based zero-crossing search algorithm on the output of a $\nabla^2 G_\sigma$ filtered image. Early experiments with this approach on typical gel images indicated that the standard LOG operator could not handle the diffuse edge character of protein spots in a robust manner.

The approach we have taken in solving this deficiency is to add a preprocessing stage to the standard LOG/zero-crossing edge detector, which we will denote by LOG-ZCED. This preprocessing stage consists of a non-linear operation added to the output of a LOG filter. The overall operation may thus be described by:

$$\nabla^2 G_{\sigma_2}(x,y) * \zeta\{\nabla^2 G_{\sigma_1}(x,y) * I_i(x,y)\} \rightarrow ZCED$$

where $\zeta\{\cdot\}$ represents a particular non-linear operation to be performed on the output of the first LOG filter, and * denotes the mathematical operation of convolution. We have examined the use of two specific non-linear operations, namely

$$\zeta\{\cdot\} \equiv |\{\nabla^2 G_{\sigma_1}(x,y) * I_i(x,y)\}| \quad (3a)$$

and

$$\zeta\{\cdot\} \equiv \{\nabla^2 G_{\sigma_1}(x,y) * I_i(x,y)\}^2 \quad (3b)$$

In practice, the second form gives slightly better overall performance and that is the form we use in our implementation. The addition of this non-linear operation achieves the desired behaviour of transforming diffuse edge boundaries into boundaries more closely resembling step-edges, which may then be extracted with a standard LOG operation. This transforming property is illustrated in Figure 2, (a) is a portion of an original gel image, and (b) is the result of applying the NLOG operator to the original image. The sequence of operations applied to the original image $I_i(x,y)$ to produce an output 'image' $\hat{I}(x,y)$ is given by

$$\hat{I}(x,y) = \nabla^2 G_{\sigma_2}(x,y) * \{\nabla G_{\sigma_1}(x,y) * I_i(x,y)\}^2 \quad (4)$$

$$I_E = ZCED\ \{\hat{I}(x,y)\} \quad (5)$$

where I_E is a binary-valued edge pixel map of the original image. The zero-crossing detection algorithm flags edge pixels in $\hat{I}(x,y)$ by applying a set of heuristics to the four nearest-neighbor pixels of a candidate edge pixel. These heuristics check for both a true zero at the candidate edge pixel location, and for sign transitions at that location which also indicate the presence of an edge pixel. Since we do not use edge linking at this step, the zero-crossing algorithm is a much simplified version of the one described by Huertas and Medioni (1986).

Spot detection and false alarm rejection

The multi-stage edge detection algorithm described above generally satisfies the high sensitivity requirements for protein spot detection in digitized gel images. However, at this high sensitivity the false alarm rate would clearly be unacceptable. The final stage detection algorithm described in this section is designed to utilize known characteristics of 'true' protein gel spots in order to provide robust false alarm rejection. Since a number of quantitative parameters describing each detected spot must be extracted and stored in a gel database, production of a high sensitivity edge map must be followed by a spot parameter extraction operation. Before describing the procedure we have developed to carry out this step, it is useful to discuss some of ways in which a given protein spot may be parameterized. We limit ourselves here to discussion of quantitation of the direct observable, i.e. measured optical density (OD) of the protein gel, in order to avoid the numerous complications introduced by attempts to quantitate in terms of absolute protein concentration. Clearly the total spot density (integrated OD) is of primary concern in gel-to-gel comparisons; in addition, we also wish to provide the ancillary parameters of area-normalized OD, and a 'total-normalized' OD, which is defined as the spot integrated OD divided by the total integrated OD

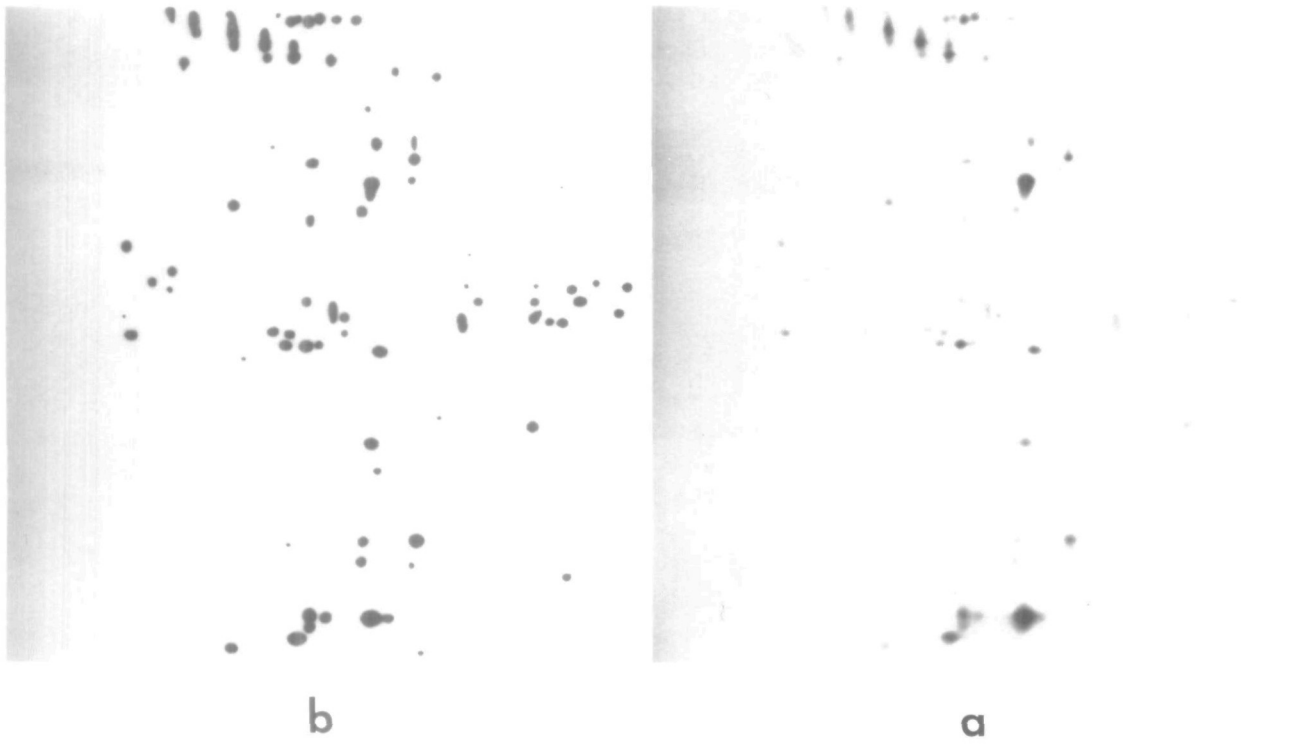

Fig. 2. Illustration of the behaviour of the NLOG operator applied to a portion of a 2DGE image of CSF proteins: (**a**) original digitized image; (**b**) image resulting from application of the NLOG operator.

of all detected spots. Both of these normalizations are useful in accommodating sample loading variabilities in the face of uncertainties regarding absolute calibration. A final intensity parameter, a smoothed estimate of the peak value within the boundary, is included to achieve a three-way background rejection criterion. Of course one must know where the spot is located within its gel, and be able to compute relative distances between spots, so the horizontal and vertical coordinates of the spot centroid become additional parameters.

In addition to parameterization in terms of optical density, one observes that protein gel spots may also be parameterized (or classified) according to shape descriptors. For example, the vertically elongated, relatively fuzzy, spot distributions frequently observed in protein gels of samples derived from body fluids of vertebrates (CSF, blood plasma, etc.), are generally associated with complex glycoproteins. Another established association is that very similar shape/intensities of either fuzzy or more distinctly bounded proteins in close proximity to each other are often clinically, biologically or genetically related to each other. A further phenomenon is that unresolved, i.e. overlapping, proteins frequently have indentations in their boundaries that can be used to flag the fact that a single detected blob is actually a multicomponent spot. Thus shape descriptors prove to be extremely useful for both intra- and inter-gel classification studies. Currently we incorporate two shape descriptors into our parameterization, both of which relate to spot elongation characteristics. The first is insensitive to the direction of elongation, and is the standard ratio of area to perimeter squared shape parameter (Duda and Hart, 1973). The second is direction sensitive, and is based on the observation that spot elongation in protein gels is generally either in the vertical or horizontal direction; the parameter used is thus simply the ratio of the vertical to horizontal dimensions of the spot, denoted as VtH ratio. This parameter is also extremely useful in providing rejection of some of the most common artefacts observed in gel images.

We have chosen the eight-direction chain code technique of Freeman (1974) for spot extraction since all the necessary spot parameters can be computed directly from the chain code representation; it is also a compact code for storage in a protein gel database. The chain coding operation is applied to the edge map produced by the operations described in the preceding section; in order to produce a valid code sequence, we demand that the boundary be closed, and that it must not have gaps larger than one pixel. This step provides a significant amount of noise rejection while still preserving likely protein spot candidates. As a matter of implementation, the chain codes are represented as linked lists using C language structures; the entire code set for a given gel image is held in an array of pointers to structures. It should be noted that this is an intermediate data structure, the final data structure being a doubly linked list of spot parameters including the spots' unique chain codes. In

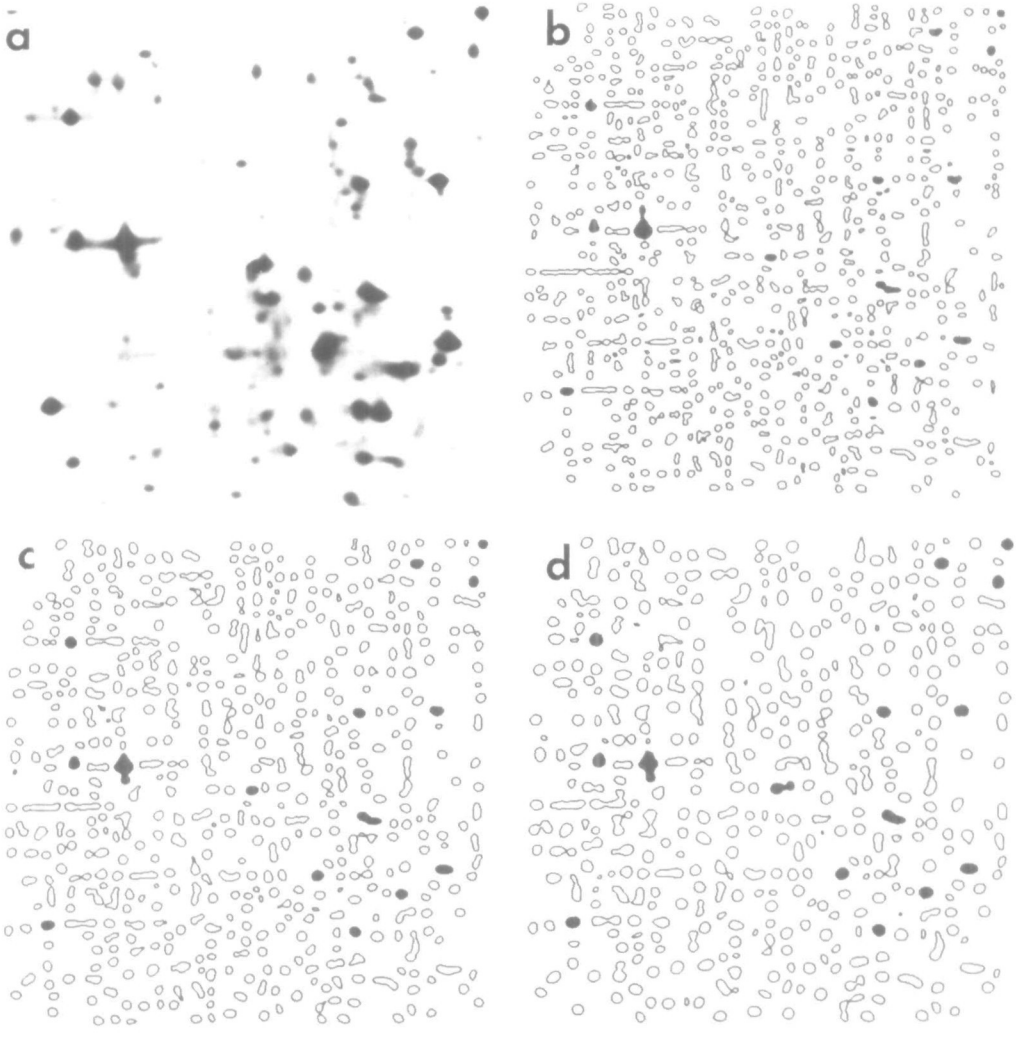

Fig. 3. Example of spot detection algorithm results when applied to a 2DGE image of proteins derived from sea urchin embryonic tissue: (**a**) original digitized image (**b**)–(**d**) spot detection results with **filt/edge** parameters set to: (5,7), (7,9) and (9,11) respectively. The most intense spots within detection edge boundaries in (b)–(d) have been filled in to aid orientation

summary, the following physical parameters are currently used to characterize detected objects which are candidate protein spots:

area_norm_od—area normalized spot OD;
centroid—spot centroid *x*- and *y*-coordinates;
concave—number of concavities in spot boundary;
gelnorm_od—total density normalized spot OD;
integrated_od—integrated spot OD;
max_od—peak spot OD;
ratio_ap2—spot area-to-perimeter squared ratio;
vh_ratio—spot vertical-to-horizontal dimension ratio;

The final step in spot detection and extraction for insertion in the gel database consists of scanning the chain code structure array to compute the relevant spot parameters from the chain code representation and the original image data. Densitometry calibration coefficients are applied at this time to obtain the actual OD values from the recorded 12-bit image data; this is also the step at which our final false alarm rejection heuristics are applied. For inclusion in the final spot list, a candidate spot must satisfy threshold criteria on **integrated_od**, **area_norm**, **max_od**, **ratio_ap2** and **concave**. As illustrated below, this multiparameter set provides robust rejection of noise- and clutter-class objects, while preserving a high probability of detection of true spots.

Implementation

The spot detection and characterization algorithm has been implemented in standard C language source code for use in the Unix operating system, and is invoked from the shell command line with two keywords, **filt** and **edge**, and an input image filename specification. The **filt** parameter specifies the spatial support (in pixels) of the non-linear LOG operator, while the **edge** parameter specifies the spatial support of the linear LOG

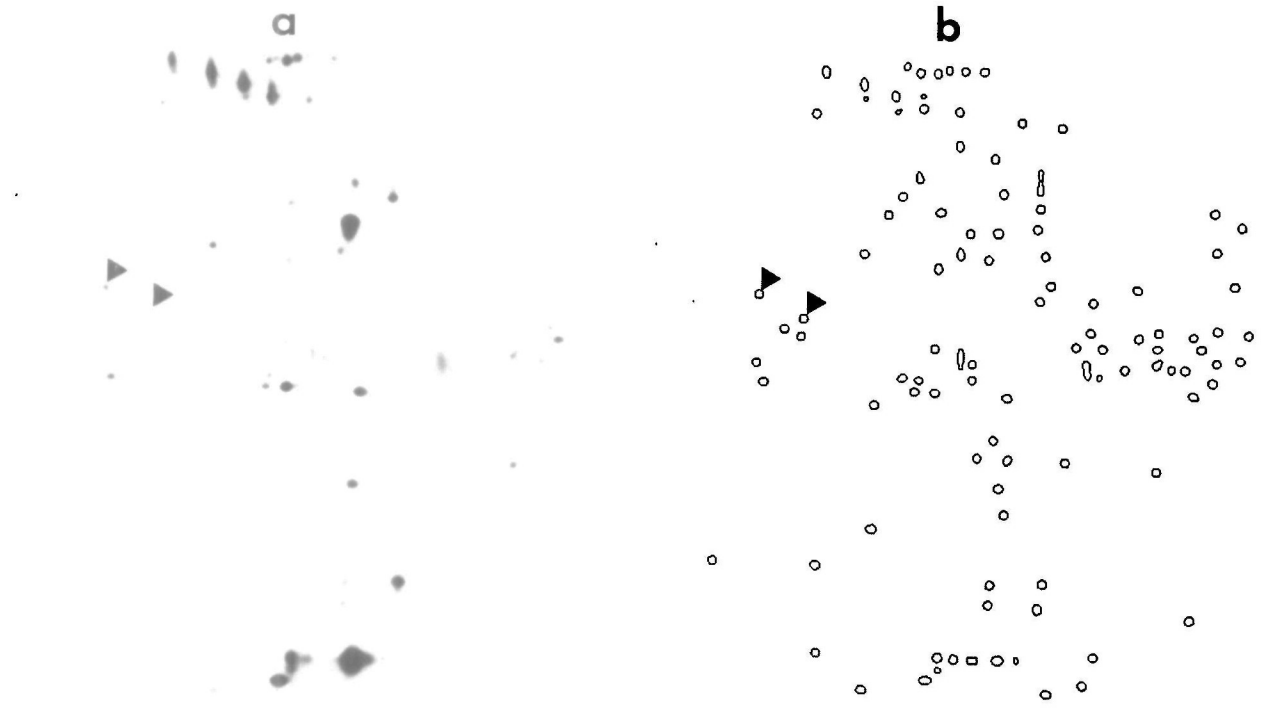

Fig. 4. Example of spot detection algorithm results when applied to a 2DGE image of proteins derived from CSF: (**a**) original digitized image; (**b**) detection results with **filt/edge** parameters set to (9,11). Arrows indicate two proteins that are diagnostic for Creutzfeld–Jacob disease.

operator. Default values of these parameters are set at **filt** = 9, and **edge** = 11. The input image file is assumed to be in raster format and of data type **short int**, with a dynamic range of 0–4095. Image files in our local image-processing software have separate ASCII 'header' files which describe the content of a specific image data file; thus there is always a header file associated with a binary image data file, e.g. **image1.hdr**, **image1.img**. The user is not required to specify an output filename, since the program automatically assigns it the name *innames***pt.spots**. All of the relevant parameters for detected protein spots, including the boundary chain code, are written to the output file as a doubly-linked list in a C-structure. Typical run-times (wall clock) for 512 × 512 pixel gel images are ~190 s for **filt** = 7, **edge** = 9 and ~230 s for **filt** = 9, **edge** = 11, on a Sun Microsystems, Sparcstation 1+; and ~120 and ~145 s respectively on a Silicon Graphics SGI240 (running one processor only). Note that runtime (for a given specification of **filt** and **edge** parameters) depends almost exclusively on image size, not on the number of detected spots, since by far the most computationally intensive part of the algorithm is in applying the edge detection operators.

Discussion

Two specific examples of applying this detection/characterization algorithm to actual 2DGE image data are shown in Figures 3 and 4, in which the data were digitized from silver-stained gels of samples derived from sea urchin embryonic tissue and from human CSF respectively. Figure 3(a) shows a 512 × 512 pixel area of the original gel data, while Figure 3(b)–(d) shows the results of applying the spot detection algorithm with **filt** and **edge** parameters set at 5 and 7, 7 and 9, and 9 and 11 respectively. For standard 16 × 20 cm gels, with 80 μm spatial sampling, spatial scale parameters of (7,9) or (9,11) are generally close to optimal both with respect to reliable detection and with respect to separability. The purpose of this figure is to give the reader some feel for the effect of selecting different **filt** and **edge** parameters; for clarity, the most intense spots within detected edges have been filled in. Figure 4(a) shows a 512 × 512 pixel area of an original silver-stained gel of human CSF from a patient with Creutzfeld–Jacob disease; two protein spots that are diagnostic for this disease state (Harrington *et al.*, 1986) are indicated by arrows. Figure 4(b) shows the results of running the detection algorithm on this data; again, the diagnostic protein spots are indicated by arrows. At the total protein loading used for this gel, the peak optical density of the darkest diagnostic spot is only 0.045 OD; for comparison, the optical density of the dark spot to the upper left of these two spots is ~0.08 OD, while the peak of the darkest detected spot in the image represents an optical density of ~2.1 OD. In fact, it has been our experience that faint spots which are not confidently visible by eye on the original gel are detected by the spot finder, and subsequently confirmed as being 'real spots' by their appearance in 5-fold total integrated intensity when a gel with 5-fold more protein sample is run.

Several other points regarding the efficacy of this algorithm are worth noting. First, its operation does not require any preprocessing either to remove or 'smooth' gel background; and so-called 'streak artefacts' do not affect its performance. Although it has been designed as a general protein spot finding tool, the use of a two-level approach, i.e. edge-finding followed by a spot validation operation, allow it to be 'tuned' for finding protein spots having very specific characteristics.

We have described a general multi-stage detection and characterization procedure for fuzzy objects, with a specific application to the analysis of protein gel images. The procedure utilizes a non-linear edge detector, based on the LOG operator, followed by a boundary following chain-coder and object classifier that performs false alarm rejection. Our experience with this approach has shown that it performs robust protein spot detection down to the level of sensitivity afforded by the 2DGE technology for protein separation, with very good false alarm rate properties. Subjectively, its performance appears to be about as good (in some cases better than) as that of visual inspection of a protein gel by a person experienced in interpreting such data. This algorithm has been implemented as a standalone program that may be run directly from the Unix command line. It has also been incorporated into our complete two-dimensional gel analysis package (GALTOOL), which we have developed and used in our laboratory over the past two years.

In addition, we have used this same algorithm with success in peak-finding and characterization in automated interpretation of two-dimensional NMR spectra of proteins in solution. The source code implementation of this algorithm may be obtained by contacting the authors via electronic mail at jerry@maxjr.hood.caltech.edu.

Acknowledgements

The authors gratefully acknowledge the assistance of Miki Yun in preparing the many protein gels, and their digitized images, which were used in the course of this investigation. The work described was supported under grant NSF-DIR-8809710, as part of the National Science Foundation STC for Molecular Biotechnology.

References

Canny,J. (1986) A computational approach to edge detection *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-8**, 679–698.

Duda,R.O. and Hart,P.E. (1973) *Pattern Classification and Scene Analysis*. John Wiley, New York.

Freeman,H. (1974) Computer processing of line-drawing images. *Comput. Surv*, **6**, 57–97.

Garrels,J.I. (1989) The QUEST system for quantitative analysis of two-dimensional gels. *J. Biol. Chem.*, **264**, 5269–5282.

Harrington,M.G., Merril,C.R., Asher,D.M. and Gajdusek,D C (1986) Abnormal proteins in the cerebrospinal fluid of patients with Creutzfeldt–Jakob disease. *New England J. Med.*, **315**, 279–283.

Huertas,A. and Medioni,G. (1986) Detection of intensity changes with subpixel accuracy using Laplacian–Gaussian masks. *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-8**, 651–664.

Marr,D. and Hildreth,E. (1980) Theory of edge detection. *Proc. Roy. Soc. London B*, **207**, 187–217.

O'Farrell (1975) ???.

O'Gorman,L. and Sanderson,A.C. (1984) The converging squares algorithm: an efficient method of locating peaks of multidimensions. *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-6**, 280–288.

O'Gorman,L. and Sanderson,A.C. (1986) Some extensions of the converging squares algorithms for image feature analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-8**, 520–524.

Skolnik,M.M., Sternberg,S.R. and Neel,J.V. (1982) Computer programs for adapting two-dimensional gels to the study of mutation. *Clin. Chem.*, **28**, 969–978.

Torre,V. and Poggio,T. (1986) On edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, **PAMI-8**, 147–163.

van Vliet,L.J., Young,I.T. and Beckers,G.L. (1989) A nonlinear Laplace operator as edge detector in noisy images. *Comput. Vision, Graphics, Image Proc.*, **45**, 167–195.

Received on January 27, 1992; accepted on July 29, 1992