

An Iterative Algorithm for Segmenting Lanes in Gel Electrophoresis Images

ALEXEI M. C. MACHADO^{1,3}, MARIO F. M. CAMPOS¹, ARI M. SIQUEIRA², OSVALDO S. F. DE CARVALHO¹

¹Computer Science Department—Universidade Federal de Minas Gerais
Caixa Postal 702, Belo Horizonte/MG, CEP 30.161-970, Brazil
alexei, mario, vado@dcc.ufmg.br

²Biochemistry Department—Universidade Federal de Minas Gerais
Caixa Postal 702, Belo Horizonte/MG, CEP 30.161-970, Brazil
siqueira@icb.ufmg.br

³Computer Science Department—Pontificia Universidade Catolica de Minas Gerais
Av. Dom Jose Gaspar, 500, Belo Horizonte/MG, CEP 30.535-610, Brazil

Abstract. In this paper we discuss the application of spatial-domain filters for solving the problem of automatic lane detection on gel electrophoresis computer images. The problem can be posed as the determination of the number, location and orientation of lanes on the image, based on the analysis of their gray-level intensities. A novel iterative filtering algorithm is proposed based on the periodicity of image projection pattern. The algorithm is compared with clustering and frequency-domain filtering methods in terms of their effectiveness and efficiency from which a trade-off solution is presented.

1 Introduction

Population Genetics is a field of Biology interested in establishing likelihood relationships among individuals whose genetic proximity enables them to be grouped together. By analyzing the DNA of a sample, it is possible to find similar genetic patterns, which may give support to the inclusion of an individual into a group of known features. The gel electrophoresis exam consists in breaking a molecule into many fragments, by the action of specific enzymes. These fragments are dispersed on a medium of polyacrylamide or agarose gel to which an electric field is applied. Each fragment has distinct electric charge and molecular weight, causing them to be displaced at different rates through the gel — smaller high-charged molecules will move faster than heavier low-charged ones. After a period of time, the process is interrupted and the gel is stained so that it becomes possible to observe where the molecules stopped (Figure 1). Each stripe in the pattern is called a band. The set of bands generated by a single sample is called a lane. The comparison between an unknown individual and already-known groups is achieved by submitting the DNA of individuals from each group to the same process. Although it cannot be guaranteed that correspondent bands from two lanes resulting from the application of the same enzyme have the same base sequence, the probability of genetic identity is considerably high, mainly if this result is reinforced by experiments with several enzymes [5].

Comparing two lanes in a gel electrophoresis image is usually a complex process as the subjectiveness of human visual perception and the factors related to the experiments may lead to different conclusions, even if

the same material is applied. An automatic analysis of the band pattern of a lane could enable the evaluation of many parameters that are usually ignored by human analysis. However, basic tasks such as the identification of lanes in a gel image, easily done by human experts, emerge as problems that may be difficult to automate.

In this paper we present a novel iterative filtering algorithm for solving the problem of detecting lanes in a gel electrophoresis image. The problem of automatic lane detection in gel electrophoresis images can be defined as the determination of the number, location and orientation of lanes on the image. Although image definition may make lane detection a difficult task, two aspects help to reduce the complexity of the problem: lanes are basically vertical clusters, arranged in a regular fashion in the gel. This, however, is not a rule. The nature of the electrical field can cause the laterally positioned lanes to bend in relation to the centrally positioned ones, in the so called smiling effect. Another aspect to be noticed is that the distance between two consecutive lanes is defined by the slot-forming template through which the sample material to be analyzed is applied. A problem arises when some of the sample slots do not receive samples, which results in missing lanes that interrupt the regularity of the gel pattern.

2 Previous Work

Lane detection can be posed as the problem of grouping bands on the gel electrophoresis image, based on their Cartesian placement. Although this definition may be suitable for clustering methodologies, the results obtained with well-known algorithms such as the Maximin,

Minimum Squared Error and Isodata has proved to be insufficient (less than 70% of accuracy)[7]. The main disadvantage of applying these clustering algorithms to the problem seems to be the setting of threshold values in the stage of image segmentation.

The problem of lane detection is also intrinsically connected to a frequency analysis question [2], as the placement of lanes describes roughly a periodic pattern. Since y -coordinates of pixels belonging to bands are not essential to the location of a lane, an useful simplification to this problem would be to reduce the dimensionality of the image by taking its projection onto the x -axis. The projection reflects the overall placement of lanes, if the high frequency pattern due to the variation of gray levels presented in the image is discarded. Parker [8] shows how this simplification can be used for the segmentation of glyphs in optical character recognition problems. Lane detection can be viewed as the problem of filtering these high-frequency elements from the projection in order to recover the pattern related to the position of lanes. The application of frequency-domain filters presented much better results than the ones obtained by clustering (from 85% to 96% of accuracy) [7]. These results, however, are very sensitive to the nature of images. The fundamental frequency related to the periodicity of lanes may be displaced with respect to the beginning and to the end of the image (phase), being interpreted as missing lanes. Missing lanes can also appear in the middle of the image as a result of absence of material. Finally, lane orientation can be distorted by the electrical field (smiling effect) or due to careless image acquisition. These problems are followed by a computational cost disadvantage: frequency filters are $O(n \log n)$ in the size of projection.

This work is part of the research effort by the Computer Science Department of the Federal University of Minas Gerais for the development of a software package for gel image analysis. The ANAGEL software supports biology scientists in all phases of gel image analysis, from the data acquisition to the comparison of patterns, combining facilities for data management and inference. Lane detection is part of the automatic band pattern extraction, which provides an intelligent user interface. Eventual mistakes in this process are acceptable, as the system offers the possibility of manual definition and deletion of lanes. We have also appraised five of the best-known existent software packages for gel analysis: Fragment Manager (Pharmacia Biotechnology), DFP-DNA Fingerprint Analysis (Biotec-Fischer), Gel Manager (CSIC - Spain), IP Lab Gel (Molecular Dynamics) and RAPDist (Australian National University). Compared to these software packages, ANAGEL is the only one that offers automatic detection of lanes from images, so we believe that this work may be an original contribution to the area of biological image recognition.

3 Lane Detection as an Image-Domain Filtering Problem

Lane detection can be viewed as a problem of determining sets of bands by filtering undesirable information directly in the pixel matrix. The use of projections in this case is even more recommended than it is for frequency-domain filtering, since they are not decomposed into periodic sinusoidal functions. Missing lanes are not a problem in this case because their absence is properly represented in the projection.

3.1 Image Acquisition and Preprocessing

Figure 1 shows an example of image acquisition by scanning, with a resolution of 125 dots per inch at 256 gray levels.

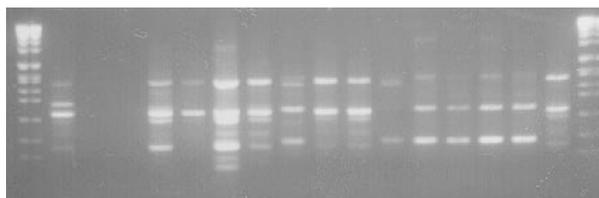


Figure 1: Original image of a gel electrophoresis of Leishmania PCR-amplified DNA samples.

Before taking the projection, the image can be filtered to reduce the noise and increase the separation of lanes. Average filtering can be done by scanning the image and computing each pixel value as an arithmetic average of the values of the pixels around it. The dimension of the inspected region will influence the degree of blurring obtained with this method. This filter is useful in reducing the impact of high-frequency noise (as salt and pepper noise) although it may also reduce the distinction of neighboring lanes. Another commonly used filter is the median filter. It computes the value of a pixel as the median of a region in the image. It reduces high-frequency noise without reducing sharply the separation of lanes. The main disadvantage of this filter is its computational cost, since it demands sorting. The last filter considered for the preprocessing phase is the minimum filter, which presented the best results for gel electrophoresis images. As the name indicates, the value of a pixel in the image is assigned the smallest value of the pixels in the region around it. This filter reduces high-frequency noise and increases the separation between the lanes, as it works eroding the objects in the image. If a mask with large dimensions is used, however, weak thin bands may disappear from the image. Figure 2 shows the result of filtering the original image depicted in Figure 1 with a mask of 3×7 , which took 0.34 seconds of CPU time. All the algorithms in this paper were implemented in C language,

compiled with the System V version of the C compiler for the SunOS 4.1 operational system and run in a Sun SparcStation 4.



Figure 2: The results of filtering the image of Figure 1 with minimum filter and mask of 3×7 .

3.2 Description

One advantage of describing the image as a projection is that there is no need for object segmentation, with the added bonus of not requiring any threshold setting. There are two main techniques of computing the projection $P(x)$ of an image I of size $n \times m$ onto the x-axis. The first, and more commonly used, takes the sum of gray levels from each column as

$$P(x) = \sum_{i=1}^n I(i, x) \quad \forall x = 1 \dots m$$

One advantage of this method is that it reduces the effect of high-frequency noise. On the other hand, lanes composed by few weak bands do not become evident and thus may be excluded during filtering phase. Low-frequency noise such as big blurs may also become too visible in the projection, since each of their pixels will contribute to the final value. This kind of noise may eventually be mistaken as lanes. An alternative to sum projection is to take the maximum value of each column as

$$P(x) = \max\{I(i, x) \quad \forall i = 1 \dots n\} \quad \forall x = 1 \dots m$$

Figure 3 and 4 show, respectively, the projection of the image from Figure 2 computed by sums and maxima.



Figure 3: The projection of the image of Figure 2 onto x-axis, using sums.

Because filtering the whole image before taking its projection is computationally expensive, an alternative to this procedure is to compute the projection first and then apply a filter to it. In most cases, the results are as good as filtering the image but the processing time is much



Figure 4: The projection of the image of Figure 2 onto x-axis, using maxima.

smaller. Another advantage of filtering the projection is that we can combine the action of different filters, so each one will act on different aspects of the problem.

3.3 Maxima Detection

After filtering, the projection presents many local maxima that may indicate the presence of lanes in the image. These maxima can be easily determined by taking the second-order derivative of the function. Depending on the image, however, filtering the projection once is not sufficient for detecting all the lanes correctly. An interesting observation to be made about the filtering process is that the number of maxima behaves as a decreasing monotonic function. This means that if we filter the projection at a time i and j and take the number n_i and n_j of maxima revealed in the respective filtering process, n_j will be at least equal to n_i , if $i < j$. An alternative method for determining the cut-off frequency is based on this overall behavior:

Iterative Algorithm for Image-Domain Filtering

1. Filter the projection with the specified filter
2. Find the maxima
3. Sort the distances between each pair of neighboring maxima and find the median
4. If there is any distance that is not an integer multiple of the median then go to step 2
5. else return.

The objective of iteration is to reduce the number of maxima generated by high frequencies in the projection by smoothing out the function. The halting condition is based on the fact that the comb used to apply the material into the agarose media has equidistant compartments. If some lane does not contain any material, the width of the blank space between the lanes in the image will simply be twice as large. This integer relation between distances and the median must of course have some tolerance. Missing lanes are an exception in gel electrophoresis, so there is low probability of the median being not the real distance between the compartments of the comb. This method proved experimentally to present superior

results than single filtering, although in some cases it may fail to converge to a satisfactory solution.

The choice of a filter in the iterative process deserves attention. The average filter is a good choice if the projection presents good separability of lanes, what can be ensured by pre-processing it with a minimum filter. Minimum filter can also be used in the iteration, but the risk of losing weak lanes increases in this case. When the projection is well-defined, with good separability of lanes, the process tends to converge quickly to an optimum result. The choice of the width value for the filter mask is also important. On one hand, if the mask is too wide, a single iteration may be sufficient to eliminate weak lanes. On the other hand, narrow masks may not be able to reduce noise, which can be mistaken as lanes.

So far, we have been concerned with the correct detection of lanes in gel images. Sometimes, however, the correct position of a lane does not match exactly the corresponding maximum detected in the filtering process. This is due to filter effect or to the appearance of the gel, which may present different concentration of material, leading to projections that are not symmetrical with respect to their centers. The rightmost lane of the image shown in Figure 1 is an example of such irregularity. If an average filter is directly applied to it, the maximum will be detected in the left extremity of the lane and this will not be acceptable for further analysis. This may also be detrimental to the iterative filtering itself, as it displaces the actual position of the lane, leading to incorrect distances and preventing convergence.

Three alternative methods for detecting maxima were implemented. We represent the filtered projection as a series of maxima and minima in the form $a_0, b_1, a_1, b_2, \dots, b_n, a_n$ where b_i is the x-coordinate of the i^{th} maximum that is preceded and succeeded by two minima, respectively denoted by a_{i-1} and a_i . The first algorithm computes the center of area between two consecutive minima:

Algorithm 1

1. Find the values of n maxima b and $n + 1$ minima a
2. For $i \leftarrow 1$ to n do
 - (a) Calculate the area between two consecutive minima $A_i = \sum_{j=a_{i-1}}^{a_i} f_j$
 - (b) Calculate the first-order momentum between two consecutive minima $m_i = \sum_{j=a_{i-1}}^{a_i} j f_j$
 - (c) $c_i \leftarrow A_i / m_i$

The computed value c_i will be the corrected x-coordinate of lane i . This method has the advantage of considering the overall distribution of material instead of only the maximum. In images that present missing lanes, however,

the minima may be still displaced, causing the computed value to fall outside the lane, as shown in Figure 5.



Figure 5: The results of maxima detection in the projection of the image shown in Figure 1, using Algorithm 1. The original projection was filtered with average filter. It can be seen that the position of the third lane is misplaced because of the absence of one lane in the gel.

Another alternative for maxima detection is presented below:

Algorithm 2

1. Find the values of n maxima b and $n + 1$ minima a
2. For $i \leftarrow 1$ to n do $c_i \leftarrow ((a_{i-1} + a_i) / 2 + b_i) / 2$

This method has the advantage of taking into account the symmetry between two minima and the maximum of each lane. It is important to consider the maximum because it is where the material is concentrated. The algorithm does not eliminate the problem of displacement in images with missing lanes, as can be seen in Figure 6.



Figure 6: The results of maxima detection in the projection of the image shown in Figure 1, using Algorithm 2. The original projection was filtered with the average filter. It can be seen that the position of the third lane (from left to right) is still misplaced.

The third method computes the center of area comprised symmetrically around the maximum:

Algorithm 3

1. Find the values of n maxima b and $n + 1$ minima a
2. For $i \leftarrow 1$ to n do
 - (a) Find the closest minimum of b_i as $d_i = \min(b_i - a_{i-1}, a_i - b_i)$
 - (b) Calculate the area above the segment from $(b_i - d_i, f(b_i - d_i))$ to $(b_i + d_i, f(b_i + d_i))$ and assign it to A_i

- (c) Find the first-order momentum of the area considered in the previous step and assign it to m_i
- (d) $c_i \leftarrow A_i/m_i$

This algorithm eliminates the problem of symmetry in images with missing lanes. Although it may take longer to compute the position of each lane, the results are much better than the other methods. An example is shown in Figure 7.



Figure 7: The results of maxima detection in the projection of the image shown in Figure 1, using Algorithm 3. The original projection was filtered with average filter. In this case, the lanes are correctly positioned in the image, despite of missing lanes. The tangent lines and the area above are also shown.

3.4 Lane Orientation

The iterative filtering algorithm aim to determine the number of lanes and their placement in the image. As seen, the problem can be simplified by reducing the image to a set of points or to a projection function. Nevertheless, the x-coordinate of each lane is not enough for practical purpose, since the analysis of the pattern requires that its main axis be determined. Sometimes, the values resulting from a detection algorithm do not represent the center of the lanes, so a correction is needed. Finally, due to careless image acquisition or to the artifacts induced by distortion of the electrical field, some lanes may be non-perpendicular to the x-axis.

The correct position and orientation of each lane can be easily achieved by the computation of moments [4]. Horn [6] shows how to compute the zero-th, first and second moments, while scanning the image. This method is amenable to parallel algorithms and presents constant asymptotic complexity.

The x-coordinate of each lane can be used to guide the computation of moments. This is done by defining a window around the perpendicular line that passes through the x-coordinate. The width of the window is the median distance between the lanes considered in the convergence criterium of the iterative filtering algorithm. The height of the window must be the height of the image, so all bands will contribute in determining the orientation. The values of the zero-th, first and second moments of each column in the image can be also computed in the projection phase and stored for further use. The center of mass (\bar{x}, \bar{y}) and the angle of orientation θ is computed as

$$\bar{x} = \frac{M_{10}}{M_{00}}, \quad \bar{y} = \frac{M_{01}}{M_{00}},$$

$$\tan 2\theta = \frac{2(M_{11} - \frac{M_{10}M_{01}}{M_{00}})}{M_{20} - M_{02} + \frac{M_{01}^2 - M_{10}^2}{M_{00}}}$$

where M_{ij} are the image moments. Figure 8 shows the final result of lane detection using the values of center of mass and orientation to draw the axis, for an electrophoresis image that present the smiling effect. The computation of these features took 0.14 seconds of CPU time.

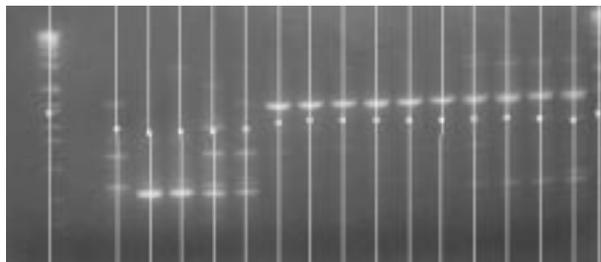


Figure 8: The final result of lane detection for an electrophoresis image that present the smiling effect.

4 Results

Table 1 summarizes the results obtained by filtering 22 gel electrophoresis images with different dimensions, with a total of 394 lanes using the iterative algorithm proposed. The projections of each image were computed by taking the sum and the maximum of their columns. Half of the experiments were conducted with the image pre-processed with a minimum filter of dimension 1×3 . Each projection was filtered once with a minimum or median filter of width 9. The iterative filtering algorithm was further applied with average filter of width 3 up to convergence. These values were selected based on the characteristics of the images. The table below shows the image identification number, the number of lane positions in each image *versus* the number of lanes that really have material, the number ε of mistakes (the number of lanes not detected plus the number of false lanes detected) and the number of iterations i for each combination of filters. Lanes were detected based on second-order derivatives. The average execution time per iteration, in each case, is shown in the bottom row of the table in seconds of CPU time, considering only the filtering phase.

All results were better for the pre-processed images, although in some cases the improvement is not worth the computational cost of filtering. It can also be observed that taking the projection with maxima is not profitable. The results of using the minimum filter in the projection pre-filtering phase are comparable to the use of median filter so it should be preferred because of the computational complexity of sorting. In some cases, the width of the

IMAGE FILTER		NO FILTER								MINIMUM							
PROJECTION		Σ				max				Σ				max			
PROJ. FILTER		min		med		min		med		min		med		min		med	
id	lanes	ϵ	i														
1	18/16	0	1	0	3	0	1	1	7	0	1	0	1	0	1	0	1
2	18/16	1	1	0	22	1	1	1	1	0	1	0	1	1	1	1	1
3	20/19	0	1	0	4	0	1	0	2	0	1	0	1	0	1	0	1
4	20/19	0	1	0	6	0	1	0	5	0	1	0	1	0	1	0	1
5	20/18	0	1	0	2	0	1	0	3	0	1	0	1	0	1	0	3
6	20/20	0	11	1	12	0	3	0	5	1	10	1	9	0	2	0	1
7	19/19	0	3	0	6	0	1	0	1	1	3	1	1	0	2	0	1
8	20/17	2	5	1	7	2	1	1	2	1	1	1	1	0	1	0	3
9	14/14	0	1	0	2	0	1	0	3	0	1	0	1	0	1	0	1
10	18/16	0	1	0	5	1	1	1	4	0	1	0	1	0	2	1	1
11	19/17	1	1	1	6	1	1	0	6	1	4	1	5	0	1	0	1
12	18/18	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
13	20/19	2	23	1	9	5	11	4	10	1	1	1	3	2	8	4	12
14	19/18	1	6	1	10	2	3	3	31	1	3	1	3	2	8	2	7
15	15/15	1	31	1	31	1	8	0	1	0	31	0	31	0	1	0	4
16	16/16	0	1	0	2	0	1	0	3	1	1	0	2	0	1	0	1
17	16/16	3	1	3	1	1	1	0	1	3	3	1	31	2	1	0	1
18	13/12	2	51	2	51	2	1	2	8	2	1	2	22	2	2	1	3
19	16/15	0	1	0	6	0	1	1	3	0	1	0	1	1	1	0	2
20	16/15	1	5	1	13	1	2	0	4	1	1	1	1	2	5	2	7
21	18/15	1	1	1	6	0	1	0	3	0	1	1	1	1	1	0	1
22	21/17	2	6	3	14	5	4	13	31	1	1	3	1	3	1	3	1
Total	392/365	17 (4%)		16 (4%)		22 (6%)		27 (7%)		14 (4%)		14 (4%)		16 (4%)		14 (4%)	
CPU time/iter.		0.009		0.010		0.011		0.010		0.010		0.009		0.010		0.010	

Table 1: The results obtained with the iterative algorithm for image-domain filtering. The first two columns show the image identification number, the number of lane positions in each image and the number of lanes that really have material. The remaining columns show the results of taking the original images (no filter) and the images with the minimum filter as the base for the projection construction. In each case, the projections were taken by sums (Σ) and maxima (*max*). Each projection was filtered once with a minimum (*min*) or median filter (*med*), before the application of the iterative process. The results are expressed in terms of the number of mistakes (ϵ) and the number of iterations (i) needed for convergence. The two last rows show the total number of mistakes, the error rate and the average execution time in seconds for each method.

mask used for pre-filtering the projection was too high, causing weak lanes to disappear before iterative filtering. Another problem has to do with lanes that are clipped in the borders of the image. In this case, the first minimum of the projection is detected after the first lane and the last minimum before the rightmost lane.

Experiments show that the methods for detecting maxima by second-order derivative and center of mass present the same error rates and speed of convergence. Nevertheless, they are important for the centering of lanes and analysis of pattern, so the approach by momentum is preferable. Considering the cost and benefits of each combination of filters, a compromise solution for the problem could be the use of raw image, sum-based projection, minimum pre-filtering and iterative filtering by average until convergence. This combination was able to detect correctly 96% of the lanes.

5 Conclusion and Future Work

We have discussed the problem of lane detection in gel electrophoresis images based on a spatial-domain filtering approach. Clustering would be the most natural approach for grouping the bands of each lane, but it involves the open problem of segmentation. An alternative method to solve the problem explores the fact that gel electrophoresis roughly describe periodic pattern. The aim of the frequency-domain filtering approach is to determine the fundamental frequency associated with the placement of lanes. The results obtained with frequency-based filters are superior than the ones obtained with clustering algorithms, but they are very sensitive to the nature of the image being considered.

Image-domain filters are broadly used for image processing due to the simplicity of algorithm implementation. The main challenge of this approach is to determine a combination of filters that is effective for solving the problem in most cases, at a compromise solution with respect to the computational cost. Many combinations were implemented. The raw image was pre-processed with average, median and minimum filter. The projection was taken based on sums and maxima of each column and processed with median and minimum filters. The iterative algorithm proposed for filtering in the frequency domain was adapted for filtering in the image domain with satisfactory results. The best compromise solution was obtained with the raw image, sum-based projection, minimum pre-filtering and average filtering in the iterative process, which also correctly detected 96% of the lanes, at a lower computational cost.

The reduction in the amount of mistaken lanes can use much improvement. If response time is reasonable, the frequency filtering approach may be combined with image-domain filter to increase the reliability of a solution. The information of distance between lanes which

was determined in the iterative filtering process can also be used to reduce the manual work done by the scientist. If an existent lane was not detected by the algorithm, the interface may provide a facility for the user to click the mouse cursor near the lane, so the system is able to discover it. With this simple procedure, the solution presented can be improved with almost no additional costs regarding to response time.

References

- [1] Ballard, D. H. & Brown, C. M. *Computer Vision*. Prentice-Hall, 1982.
- [2] Bracewell, R. N. *The Fourier Transform and Its Applications*. McGraw-Hill, 1965.
- [3] Castleman, Kenneth R. *Digital Image Processing*. Prentice Hall, 1979.
- [4] Gonzalez, R. C. *Digital Image Processing*. Addison-Wesley, 1977.
- [5] Hoelzel, A. & Dover, G. *Molecular Genetic Ecology*. IRL Press, Oxford, 1991.
- [6] Horn, B. *Robot Vision*. MIT Press, 1986.
- [7] Machado, A. & Siqueira, A. & Carvalho, O. & Campos, M. *Automatic Lane Detection in Gel Electrophoresis Images*. Technical report DCC-013/97, Universidade Federal de Minas Gerais, 1997.
- [8] Parker, J. *Algorithms for Image Processing and Computer Vision*. John Wiley & Sons, 1997.
- [9] Rickwood, D. & Hames, B. *Gel Electrophoresis of Nucleic Acids: a practical approach*. IRL Press, Oxford, 1982.
- [10] Rosenfeld, A. & Kak, A. *Digital Picture Processing*. Academic Press, 1976.