Structural Matching of 2D Electrophoresis Gels using Graph Models

Alexandre Noma[1*], Alvaro Pardo[2†], Roberto M. Cesar-Jr[1*]
[1] IME-USP, Department of Computer Science, University of São Paulo, Brazil
[2] DIE, Faculty of Engineering and Technologies, Catholic University of Uruguay

Abstract

2D electrophoresis is a well known method for protein separation which is extremely useful in the field of proteomics. Each spot in the image represents a protein accumulation and the goal is to perform a differential analysis between pairs of images to study changes in protein content. It is thus necessary to register two images by finding spot correspondences. Although it may seem a simple task, generally, the manual processing of this kind of images is very cumbersome. The complete task of individual spot matching and gel registration is a complex and time consuming process when strong variations between corresponding sets of spots are expected. Besides, because an one-to-one mapping is expected between the two images, missing spots there may exist on both images (i.e. spots without correspondence). In order to solve this problem, this paper proposes a new distance together with a correspondence estimation algorithm based on graph matching which takes into account the structural information between the detected spots. Each image is represented by a graph and the task is to find an isomorphism between subgraphs. Successful experimental results using real data are presented, including a comparative performance evaluation.

1. Introduction

In this paper, the problem of 2D electrophoresis gel matching is addressed. Two-dimensional electrophoresis is a well known method for protein separation which is extremely useful in the field of proteomics. The basic idea is to separate proteins contained in a sample using two independent properties such as isoelectric point and mass. An example of images that are obtained is given in Figure 4. Each spot in the image represents a protein accumulation and its size depends on the amount of protein present in the sample. The grayscale on top of each image is placed to allow grayscale calibration. Although it may seem a simple task, the manual processing of this kind of images is very cumbersome. Furthermore, since gel electrophoresis is generally used to compare samples, several pairs of images must be compared during a single experiment. For this kind of differential analysis, it is necessary to register two images by finding spot correspondences (Figure 1). One of the reasons for the popularity of 2D gel electrophoresis is its simplicity. As a counterpart, the experimental setting and the materials used do not allow a highly controlled experiment. This means that strong variations between corresponding sets of spots are in general expected.

All these elements show that, although 2D gel images may seem simple, the complete task of individual spot matching and gel registration is a complex and time consuming process.

Most of the existing methods for gel matching does not take into account the structural information between the spots to obtain the correspondence. Instead, they usually start by extracting point features to represent the spots which are then used for point and gel matching. In some cases these point features can be used to establish point correspondence before obtaining the complete gel matching [8]. In [8] the authors present a method to match sets of 2D points using an iterative algorithm that combines point correspondence and transform estimation. In order to establish point correspondences they separated the procedure into distance computation and correspondence estimation. Finally, based on the point correspondence they estimate a transformation between both sets of 2D points. These two steps may be iterated to refine the results. The authors present a detailed evaluation for different distances between points and different point correspondence estimations. Given two sets of 2D points $x = \{x_1, ..., x_N\}$ and $y = \{y_1, ..., y_M\}$, and a distance $d_{ij} = d(x_i, y_j)$, the idea is to compute the correspondence between points. They evaluate the Euclidean distance and the Shape Context distance [4] for point matching. Regarding the correspondence estimation they propose several methods: Closest Point, k-Closest Points, Bi-Partite Graph Matching and other graph based ones.

The novelties in this article are (1) the introduction of a new distance that includes structural information in order to solve the gel matching problem, together with (2) an adaptation of the correspondence estimation algorithm based on

* E-mails: {noma, cesar}@ime.usp.br
† E-mail: apardo@ucu.edu.uy

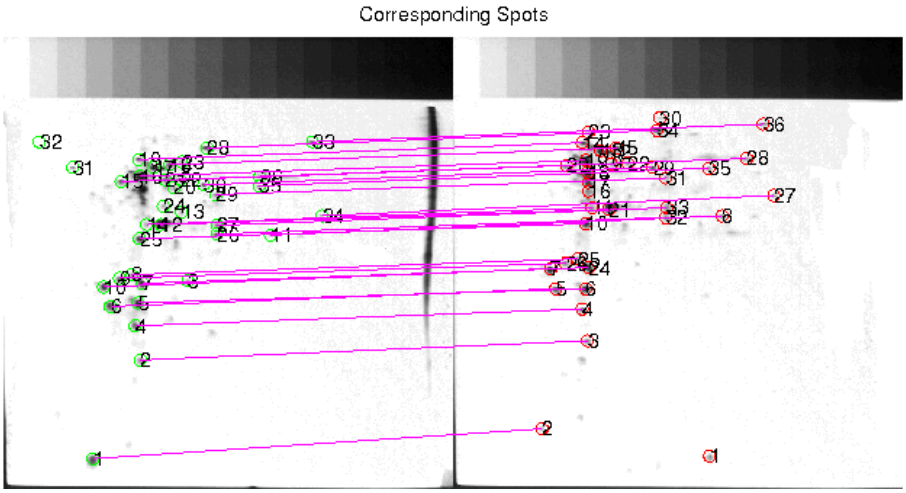

Figure 1. An obtained correspondence for a pair of gel images.

previous work [7] in order to obtain an isomorphism. Existing solutions for 2D gel registration spot matching are based on distances between pair of points (one from each image). In our case, the proposed distance reflects structural information of both images using corresponding vectors, as well as local shape information using Shape Context. Each input image is represented by a graph, where each point is represented by a vertex. Directed edges are created between vertices when there is a structural relation between those vertices (see Figure 3).

The original algorithm [7] calculates an homomorphism in order to solve a segmentation problem based on watershed basins. In our case, an isomorphism between the two graphs is necessary in order to find a correspondence between the two input gel images. In this case, a post-processing is performed and the proposed greedy algorithm minimizes:

$$E = \alpha \sum_{vertices} d_{SC} + (1 - \alpha) \sum_{edges} d_S.$$

The first term d_{SC} represents the Shape Context distance and compares pairs of vertices representing the corresponding points. The second term d_S consists of the Structural distance which takes into account the 'edge costs'. Both terms are balanced by a parameter α, which consists of a real number between 0 and 1.

The remaining of this paper is organized as follows. It starts, in Section 2, with a brief description about the initial step: the detection of the spots in each input gel image. The shape and structural distances are given in Sections 3 and 4, respectively. In Section 5 there is a detailed description about the proposed algorithm, followed by some experimental results in Section 6. Finally, the conclusions and observations are given in Section 7.

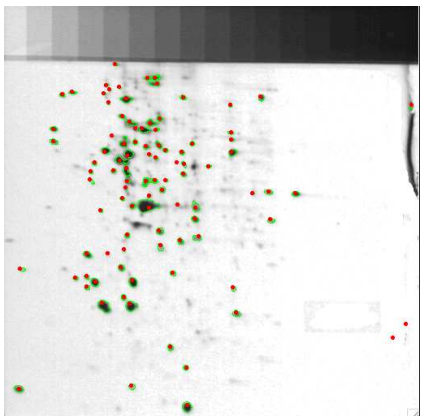

Figure 2. Detection of spots.

2. Spot Detection

As mentioned in the Introduction, matching two gel images should be based on invariant features present on them, such as points representing each spot. The algorithm proposed in [3] is used for the detection of these points. This algorithm is based on the detection of meaningful spots where meaningfulness is determined by its contrast and shape. The point descriptor of the spots is the darkest point inside it, i.e. the peak of protein concentration. Figure 2 illustrates the result of the meaningful boundaries detection algorithm applied to a real pair of gel images.

3. Spot Distance using Shape Context

The best methods for spot matching are based on point-matching techniques. In this paper, the Shape Context (SC) [4] metric is adopted, inspired by [8] where this metric was

applied to gel images. The idea behind SC is to describe each point (spot) with the distribution of points on its neighborhood. Using a set of bins in polar coordinates, the number of points in each bin is computed to obtain a 2D histogram in polar coordinates. The normalized histogram at point i is denoted as $h_i(k)$, where the index k identifies the bin. Given this metric we can compute the distance between the SC of two points i and j using the χ^2 distance:

$$d_{\chi^2}(SC_i, SC_j) = \frac{1}{2}\sum_k \frac{[h_i(k) - h_j(k)]^2}{h_i(k) + h_j(k)} \qquad (1)$$

When comparing two SC, small discrepancies between corresponding points may exist. These discrepancies may have different sources. First, there are genuine differences due to the appearance or missing spots. Second, there is the possibility of misdetection or errors in spot detection. Third, there may also be gel deformation. The algorithm proposed in [3] proposes kernel estimation of the histogram to deal with these discrepancies. This modification of SC showed robustness and better generalization capabilities.

At the end of this step of spot matching we obtain a matrix C where each entry $C_{ij} = d(SC_i, SC_j)$. Therefore, for each spot in one image we obtain the similarity with each spot in the other one.

4. Structural Distance

Besides Shape Context, structural information is used in the proposed algorithm to help the task of obtaining a good solution for the 2D gel registration spot matching.

First, the detected spots are represented as 2D points, as shown in Figure 3(a). In order to evaluate the structural distance, each input image is represented by a graph, where each 2D point is assigned to a vertex, and edges are created in order to represent structural relations between vertices. The proposed method is inspired by the graph matching approach for image segmentation described in [5, 6, 7].

More specifically, one of the two input gel images is chosen to be the model (and the other is referred simply as input). The model graph contains the structural information, as shown on the left of Figure 3(b), being denoted by $G_m = (V_m, E_m)$, where V_m is the set of vertices and E_m the set of edges. Similar notation is used for the input graph G_i. The task is to associate each input vertex v_i to a model vertex v_m, where each association is denoted by the pair of vertices (v_i, v_m), $v_i \in Vi$ and $v_m \in V_m$. Because an one-to-one mapping is expected between the two graphs, missing spots there may exist on both images (i.e. spots without correspondence), and the problem reduces to finding an isomorphism between two subgraphs, a subgraph of G_m and a subgraph of G_i.

When considering structural information, each $v_i \in V_i$ tends to be associated to the nearest $v_m \in V_m$ as shown on the right of Figure 3(b). Due to detection errors, outliers may exist. There is no correspondence to such outlier vertices. Hence, the problem of gel registration requires that the solution represents a bijection between subsets of V_i and V_m. The original algorithm [7] associates each $v_i \in V_i$ to exactly one $v_m \in V_m$, but the reversal is not always true. Eventually, distant v_i's could be assigned to a same v_m but with higher costs. In this case, for each $v_m \in V_m$, it is enough to keep the cheapest pair (v_i, v_m), $v_i \in V_i$, in the solution and to discard the pairs with 'repeated' v_m. This post-processing guarantees the expected one-to-one mapping.

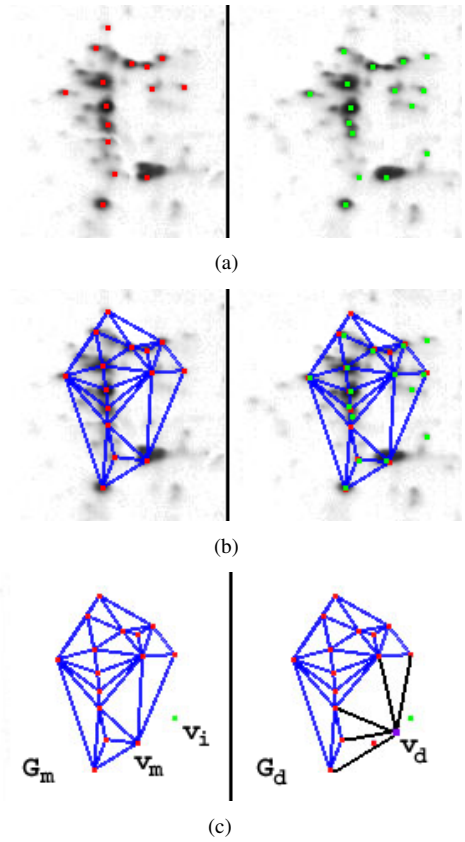

Figure 3. (a) Detected spots. (b) The structural information extracted from the left image superposed to the right image. (c) The deformation graph G_d due to a pair (v_i, v_m).

The proposed algorithm is based on a greedy strategy and is described in detail in the next Section. For each $v_i \in V_i$, it examines each candidate pair (v_i, v_m), $v_m \in V_m$, and includes the cheapest pair in the solution.

In order to calculate the structural distance related to each candidate (v_i, v_m), we use an auxiliary structure, called deformation graph G_d, which represents the local structure deformation caused by (v_i, v_m) on

the model G_m [7]. Figure 3(c) illustrates this idea. For a given G_m and a candidate pair (v_i, v_m), the corresponding deformation graph G_d is shown, where v_d is the deformation vertex obtained by taking the average of the coordinates of v_i and v_m. Note that the deformed edges $e_d \in E_d$ which are actually (possibly) deformed (compared to the model G_m) are those with an end at v_d.

The structural distance is defined as:

$$\begin{aligned} d_S(e_d, e_m) = \quad & \delta d_A(\nu(e_d), \nu(e_m)) \\ & +(1-\delta) d_M(\nu(e_d), \nu(e_m)) \end{aligned} \tag{2}$$

where e_d is a deformed edge and e_m the corresponding model edge. These edges are considered as directed edges and their endpoints are vertices representing bidimensional points. In this case, their corresponding vectors are taken into account, being denoted as $\nu(e_d)$ and $\nu(e_m)$, respectively.

The structural distance consists of two terms: angular (d_A) and modulus (d_M) distances; $d_A(\nu(e_d), \nu(e_m)) = \frac{|\cos\theta - 1|}{2}$, where θ is the angle between the two vectors; $d_M(\nu(e_d), \nu(e_m)) = \frac{||\nu(e_d)| - |\nu(e_m)||}{d_{\max}}$, where $d_{\max}$ is a normalization constant, and $|\nu(e_d)|$ denotes the vector modulus of $\nu(e_d)$ (same for $\nu(e_m)$).

The angular and modulus distances are weighted by a parameter δ, which is a real number between 0 and 1.

5. Spot Matching

The Shape Context distance defined by Equation 1 together with the structural distance in Equation 2 are used to guide the matching process to find potential correspondence candidates. The goal of the spot matching step is to find a unique correspondent for each spot. Since natural differences and detection errors are expected, some of the spots may not be matched.

The proposed algorithm is iterative and examines each input vertex $v_i \in V_i$ at a time. For each v_i, the algorithm chooses the cheapest $v_m \in V_m$ and includes the corresponding pair (v_i, v_m) in the solution. The cost of each candidate pair (v_i, v_m) is evaluated by the following equation:

$$\begin{aligned} c(v_i, v_m) = \quad & \alpha d_{SC}(SC(v_i), SC(v_m)) \\ & +(1-\alpha)\tfrac{1}{n_e} \textstyle\sum d_S(\nu(e_d), \nu(e_m)) \end{aligned} \tag{3}$$

which takes into account the shape context and the structural distances: $d_{SC} = d_{\chi^2}$ (Equation 1) and d_S (Equation 2). The first term is the SC distance between the two points represented by v_i and v_m, using the χ^2 distance. The second term represents an average of n_e deformed edges considered in the computation of the structural distance.

In order to compute a correspondence, the structural information plays an important role. It drastically restricts the number of possibilities and permits a rapid convergence to good solutions. The final solution is a set P of pairs (v_i, v_m), $v_i \in V_i$, $v_m \in V_m$, which includes only the spots having correspondence. The possibly outliers and spots without correspondence are discarded from the solution during the post-processing step. For each $v_m \in V_m$, the algorithm evaluates the cost of each pair $(v_i, v_m) \in P$, $v_i \in V_i$, and keeps only the cheapest pair (v_i', v_m). All the remaining pairs (v_i, v_m), $v_i \neq v_i'$ are removed from P.

A pseudo-code of the proposed algorithm is presented below:

- Input: G_i and G_m
- Output: Set P of pairs (v_i, v_m) representing the graph matching between G_i and G_m.

$P = \emptyset$
for each vertex $v_i \in V_i$ **do**
  $c_{\text{min}} \leftarrow \infty$
  $v_{\text{min}} \leftarrow$ NULL
  for each vertex $v_m \in V_m$ **do**
    $c \leftarrow c(v_i, v_m)$
    if $c < c_{\text{min}}$ **then**
      $c_{\text{min}} \leftarrow c$
      $v_{\text{min}} \leftarrow v_m$
    end if
  end for
  $P \leftarrow P \cup \{(v_i, v_{\text{min}})\}$
end for
Post-processing of P: for each v_m, keep only the cheapest pair (v_i, v_m) in P.
return P

---

6. Experimental Results

In this Section, the benefits of the proposed algorithm are compared to the Bipartite Graph Matching (BGM) [4] algorithm, which is one of the methods proposed in the literature [8]. For the ground truth data, images from [1] were processed with the method proposed in [3] and those manually verified.

Given a matrix of similarity between spots, C_{ij}, the idea of BGM is to find the optimal assignments which minimize the total cost of matching:

$$\min_{P_{ij}} \sum_{ij} C_{ij} P_{ij} \tag{4}$$

where P_{ij} is a permutation matrix which encodes the matching. In order to reject the outlier spots, a set of virtual spots with cost ε is included for rejection purposes. This is done by using an expanded matrix $C_\varepsilon = [C\ \varepsilon]$. The best results are defined by selecting the parameter ε^* which minimizes the number of errors.

In order to test the robustness of the proposed algorithm, the evaluation was divided into two parts. First, only shape

information is affected when some of the spots are artificially removed from both images and thus increasing the number of outliers. In the second experiment, both structural and shape information is affected by gaussian noise on the points coordinates to increase the nonrigid transformations present on both images. Results from both experiments using the pair of gels in Figure 4 are shown.

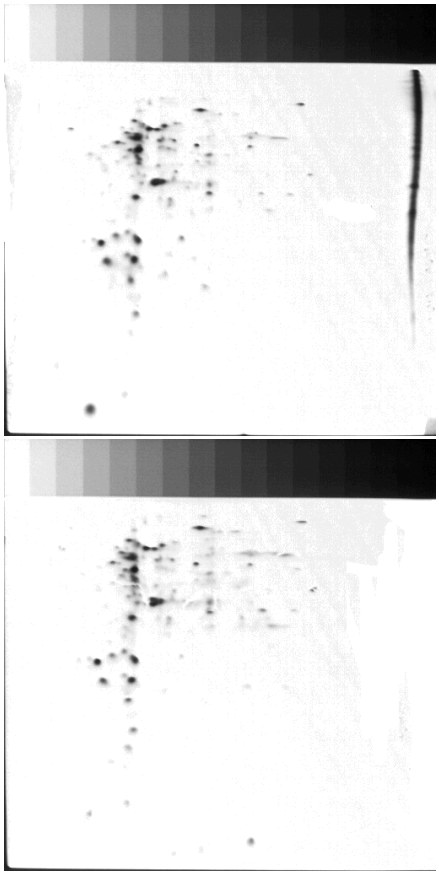

Figure 4. An example of a pair of gel images.

6.1. Experiment 1

In this first experiment, each pair of gel images were subjected to a degradation of the shape information when some spots are artificially removed from each image. This simulates the possible errors during detection and the natural differences between corresponding spots[1].

Subsets of 70, 80 and 90% of the original points were used to compare the proposed matching algorithm (which combines shape and structural information) against BGM. For each percentage, 100 trials were considered. For each artificially obtained pair, the amount of errors in the correspondence was computed (according to a ground truth) for different values of α and ε.

The parameter α in Equation 3 controls the balance between shape and structural information; for $\alpha = 0$ there is only structural information and for $\alpha = 1$ the algorithm uses only shape information. For all tests, the proposed algorithm used $\delta = 0.5$ in order to give the same importance to both terms in Equation 2. The mean and standard deviation corresponding to the best results obtained from both algorithms on the pair in Figure 4 are illustrated in Table 1, and the complete behaviour for this experiment is presented in Figures 5 and 6.

	Prop.		BGM		Prop. ($\alpha = 1$)	
	Mean	Std.	Mean	Std.	Mean	Std.
70%	**6.17**	**2.47**	7.22	2.80	8.65	3.09
80%	**5.39**	**2.29**	7.25	2.80	8.10	3.07
90%	**3.89**	**1.95**	5.67	2.80	6.19	2.80

Table 1. Best results from both algorithms. In all cases when structural information is considered, the proposed method outperforms the BGM algorithm.

From this experiment the following conclusions may be drawn. First, the inclusion of structural information improves the results (between 40 and 70%). If only Shape Context is used (letting $\alpha = 1$) then the BGM performance is slightly better, thus structural information is necessary in order to get a better performance. Second, shape information describes the local configuration of points and is crucial to decrease the matching errors. Finally, there exists an optimum $\alpha = 0.4$ which is the same for the three tested scenarios, and the error rate is stable for α in $[0.2, 0.6]$. These facts corroborate the method robustness.

6.2. Experiment 2

In this experiment, in order to increase the nonrigid transformations between points, gaussian noise was added to the points coordinates from both images according to three different standard deviations. Each image from the pair is subjected to a perturbation similar to the one illustrated in Figure 9. For each standard deviation, 100 sample pairs were generated.

The mean and standard deviation corresponding to the best results obtained from both algorithms on the pair in Figure 4 are illustrated in Table 2, and the complete behaviour for this experiment is presented in Figures 7 and 8.

As illustrated in Table 2, the proposed algorithm outperforms BGM in all cases when strutural information is con-

1 This experiment is equivalent to have an artificially generated set of corresponding image pairs.

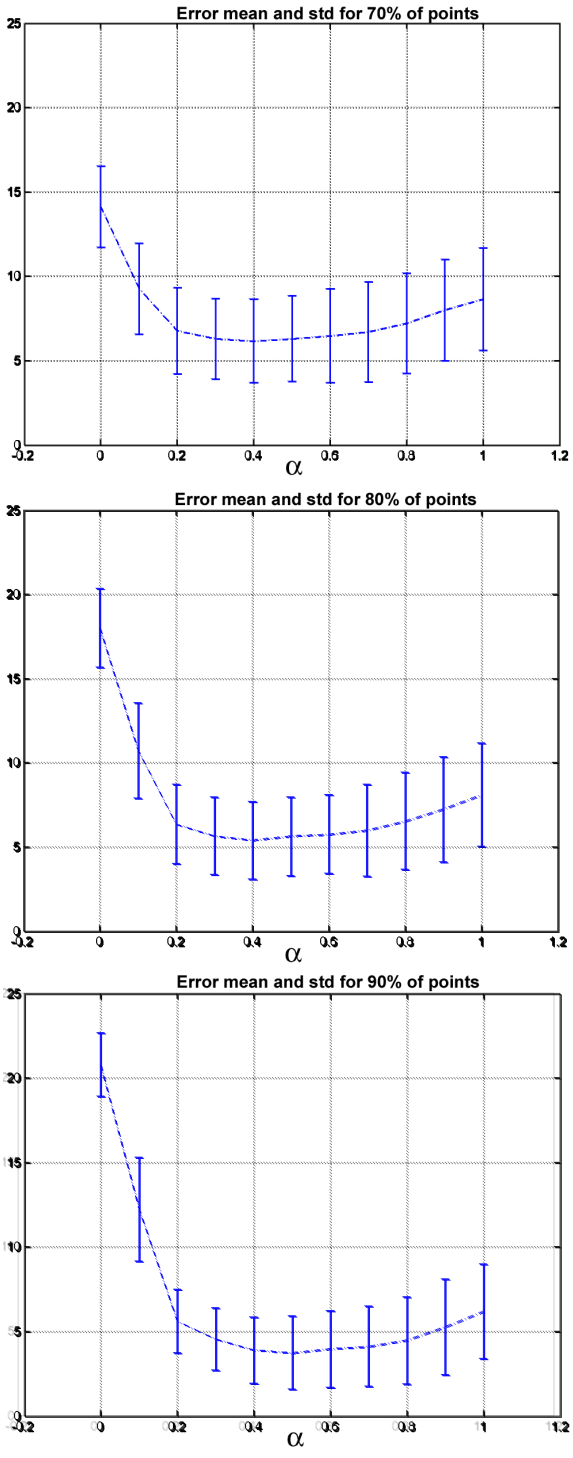

Figure 5. Experiment 1: The amount of errors obtained by the proposed algorithm for different sampling percentages (70, 80 and 90%) and for different values of α.

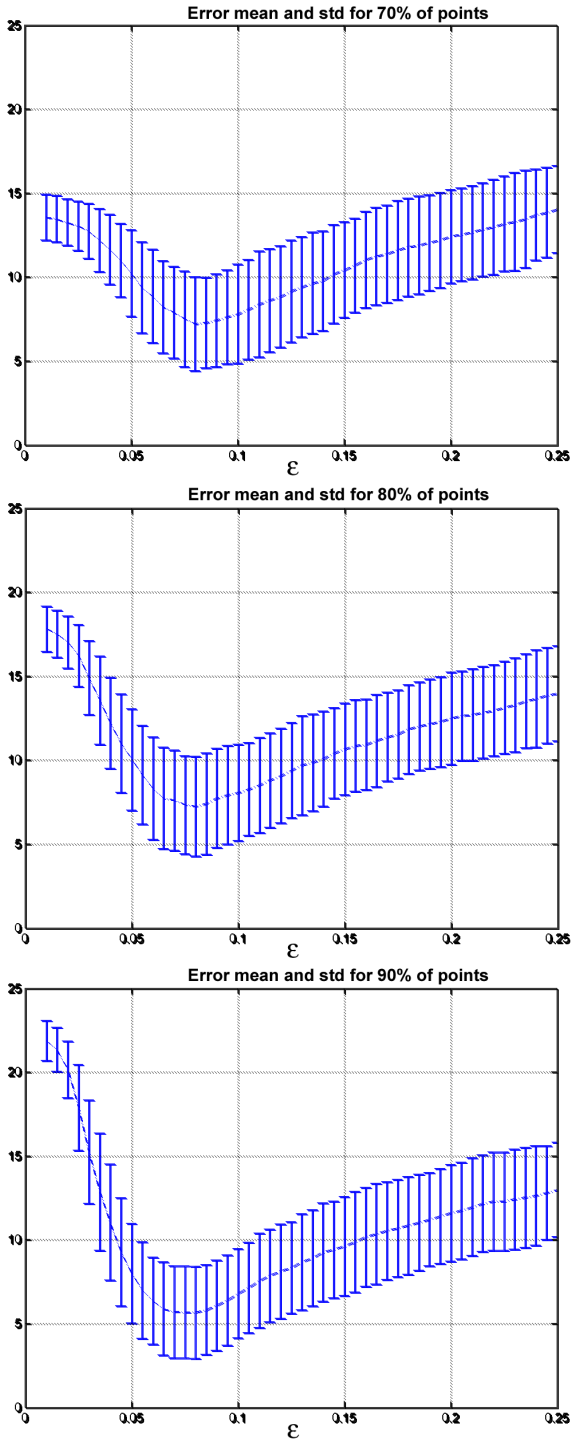

Figure 6. Experiment 1: The amount of errors obtained by BGM for different sampling percentages (70, 80 and 90%) and for different values of ε.

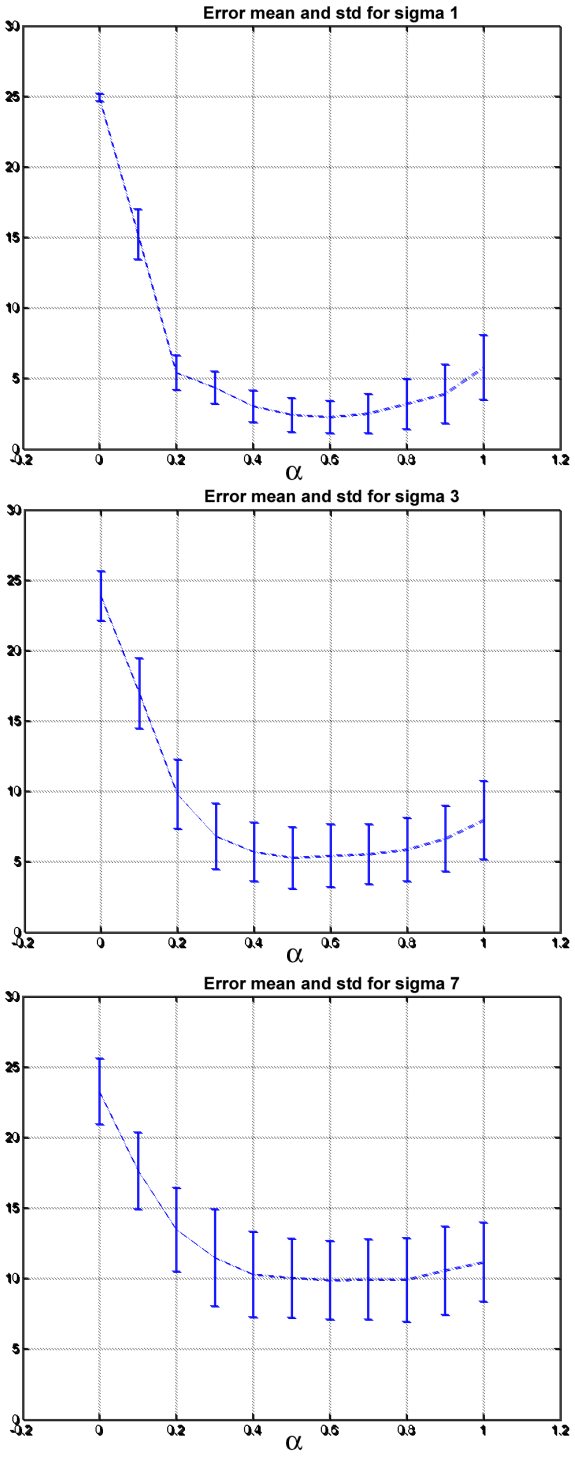

Figure 7. Experiment 2: The amount of errors obtained by the proposed algorithm for different noise variance on the points coordinates and for different values of α.

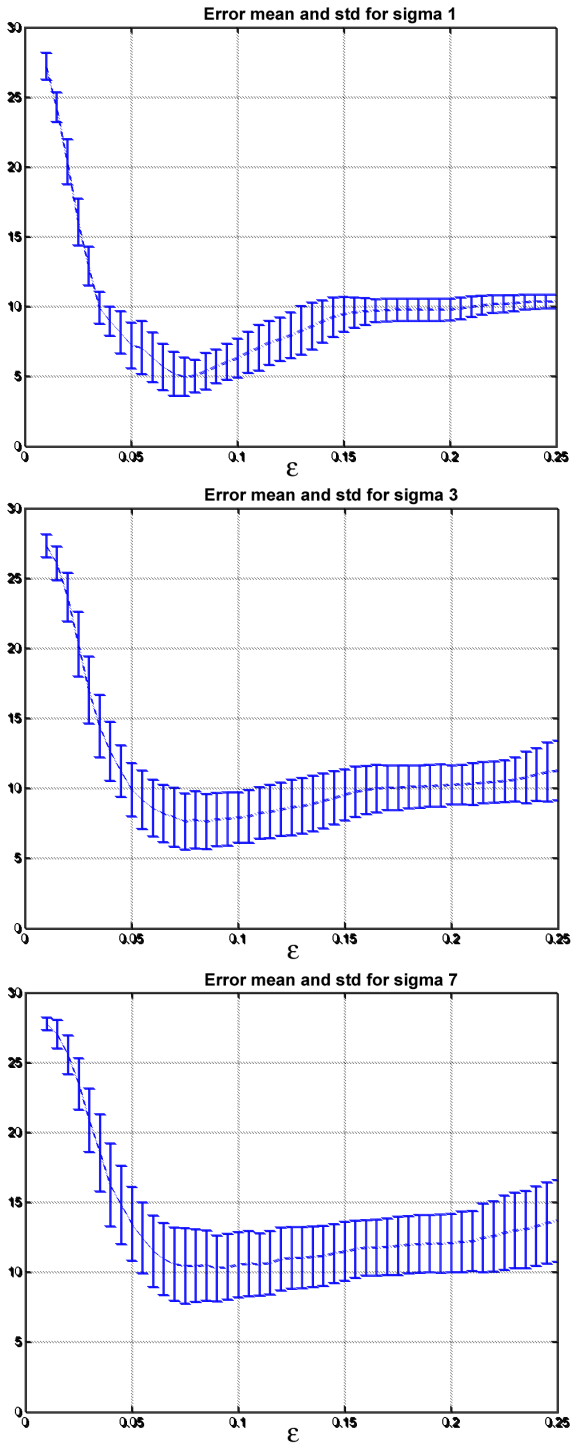

Figure 8. Experiment 2: The amount of errors obtained by BGM for different noise variance on the points coordinates and for different values of ε.

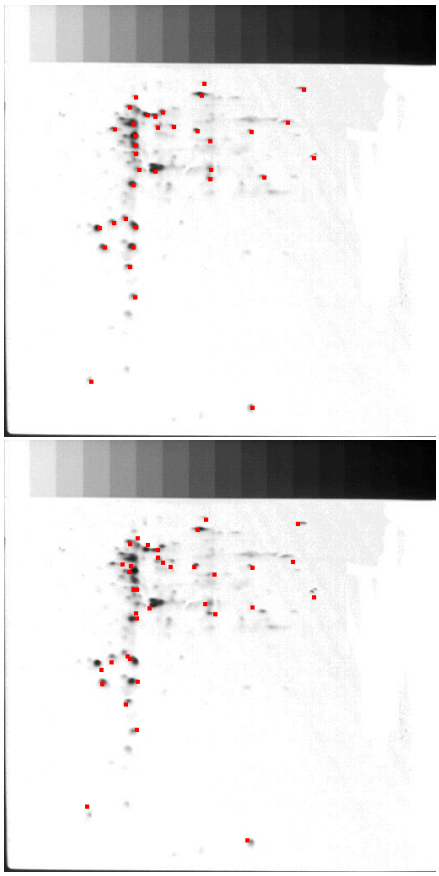

Figure 9. The original points (on top) and the corresponding points after the addition of a gaussian noise with $\sigma = 7$ (on bottom).

sidered. This corroborates to the robustness of the structural information.

7. Conclusions

In this work we presented a method for structural matching of 2D electrophoresis gels using graph models. The proposed algorithm is based on a greedy algorithm which minimizes a given energy to obtain a matching. The energy function represents a distance for spot comparison, which takes into account structural and shape information.

The proposed algorithm was tested against the Bipartite Graph Matching (BGM) [4] algorithm using a real pair obtained from [1]. The test was divided into two experiments. In the first one, only shape information is affected by artificially removing some spots from both images, thus increasing the number of possible outliers. The second experiment tested the robustness of the structural information by introducing random perturbations on the points coordinates, thus increasing the nonrigid transformations between points.

	Prop.		BGM		Prop. ($\alpha = 1$)	
	Mean	Std.	Mean	Std.	Mean	Std.
$\sigma = 1$	**2.27**	**1.15**	4.98	1.37	5.8	2.28
$\sigma = 3$	**5.28**	2.19	7.62	**1.96**	7.95	2.78
$\sigma = 7$	**9.87**	2.78	10.28	**2.35**	11.17	2.80

Table 2. Best results from both algorithms. Even in the presence of a perturbation in the structural information, the proposed method needs this information to outperform BGM.

The method has been successfully applied to different real image pairs, including higher complexity pairs, which means larger number of spots and larger deformation between images. As expected, the error rate increases for more complex pairs (see [2] for more results).

The results of both experiments confirmed the need and the robustness of structural information in order to obtain better matchings. When comparing the best results from the proposed algorithm and BGM, the proposed method presented better results in all cases when structural information is considered.

Finally, the simplicity of the new proposed method allows extremely fast implementations, and it can be easily integrated in a graphical interface for user interaction.

Acknowledgements: The authors are grateful to FAPESP, CNPq and CAPES.

References

[1] http://www.lecb.ncifcrf.gov/2dgeldatasets/.

[2] http://www.vision.ime.usp.br/∼noma/gels1e/.

[3] A. Almansa, M. Gerschuni, A. Pardo, and J. Preciozzi. Processing of 2d electrophoresis gels. *1st International Workshop on Computer Vision Applications for Developing Regions in Conjunction with ICCV*, 2007. Rio de Janeiro, Brazil.

[4] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):509–522, April 2002.

[5] R. M. Cesar-Jr., E. Bengoetxea, I. Bloch, and P. Larrañaga. Inexact graph matching for model-based recognition: evaluation and comparison of optimization algorithms. *Pattern Recognition*, 38(11):2099–2113, 2005.

[6] L. A. Consularo, R. M. Cesar-Jr., and I. Bloch. Structural image segmentation with interactive model generation. In *Proc. IEEE International Conference on Image Processing (ICIP-07)*. IEEE, Piscataway, NJ, 2007.

[7] A. Noma, A. Graciano, L. Consularo, R. Cesar-Jr, and I. Bloch. A new algorithm for interactive structural image segmentation. arXiv:0805.1854v2 [cs.CV], 2008.

[8] M. Rogers and M. Graham. Robust and accurate registration of 2-d electrophoresis gels using point matching. *IEEE Transactions on Image Processing*, 16(3):624–635, March 2007.