Structural Matching of 2D Electrophoresis Gels using Deformed Graphs

Alexandre Noma\textsuperscript{a}, Alvaro Pardo\textsuperscript{b}, Roberto M. Cesar-Jr\textsuperscript{a},

\textsuperscript{a}Institute of Mathematics and Statistics - IME
- University of São Paulo - USP, Rua do Matão, 1010, 05508-090, São Paulo, Brazil,
alex.noma@gmail.com  cesar@ime.usp.br

\textsuperscript{b}Department of Electrical Engineering - Faculty of Engineering and Technologies
- Catholic University of Uruguay, Av. 8 de octubre 2801 CP11600, Montevideo, Uruguay
apardo@ucu.edu.uy

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, especially when strong variations between corresponding sets of spots are expected (e.g. strong non-linear deformations and outliers). In order to solve this problem, this paper proposes a new quadratic assignment formulation 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 a maximum common subgraph. Successful experimental results using real data are presented, including an extensive comparative performance evaluation with ground-truth data.
1 Introduction

In this paper, the problem of 2D electrophoresis gel matching is addressed. Two-dimensional electrophoresis (2DE) 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 1 together with the corresponding spot matching.

Each spot in the image represents a protein accumulation and its size depends on the amount of protein present in the sample. A grayscale is placed on the top of each image 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 2DE is its simplicity. As a counterpart, the experimental setting and the materials used do not allow a highly controlled experiment. This means that, in general, strong variations between corresponding sets of spots are expected. All these elements show that, although 2DE gel images may seem simple, the complete task of individual spot matching and gel registration is a complex and time consuming process.

In the present work, we extend the algorithm described in (Noma et al. 2008) for the 2DE gel matching problem. For complex pairs of gels, in order to achieve translation invariance, we apply a strategy based on iterative closest point (ICP) (Besl and McKay 1992). In the experiments, we illustrate more complex cases than those shown in (Noma et al. 2008), comparing the extended approach against two well-known methods described in the literature, the graduated assignment algorithm, described in (Chui and Rangarajan 2000), which uses the Euclidean distance, and the bipartite graph matching combined with shape context, proposed in (Belongie et al. 2002), which is well-known in the literature to be very robust to outliers. Both methods are combined with transform estimation (e.g. thin-plate splines) in order to solve the point matching problem. While these two rely on transform estimation, our approach uses deformed graphs to constrain the structural relations between points. All of them are closely related to the ICP, which is attractive because of its simplicity and good convergence properties.

The proposed method is asymptotically faster than the ones proposed in the literature and can interact with the user in order to find the correspondences for complex pairs of gels, taking advantage of previous knowledge provided by the user by selecting or validating a couple of corresponding spots.

The remainder of this paper follows in Section 2 with a brief overview of related
works. Section 3 formulates the 2DE gel matching task as a quadratic assignment problem via graph matching, describing each term of the cost function and the optimization algorithm based on deformed graphs. Experimental results are illustrated in Section 4. Finally, the conclusions and final observations are given in Section 5.

2 Background

In this section, we describe the main classes of methods related to the 2DE gel matching task. This overview does not aim to be exhaustive, but just to exhibit the main trends in current research and to highlight some open issues we would like to address in this paper.

The 2DE gel matching problem is closely related to the 2D point pattern matching problem, in which there is an extensive literature with sophisticated methods, ranging from \(O(n^6)\) complexity to \(O(n^2)\) algorithms. As pointed out in (Rogers and Graham 2007), in practice 3000–4000 spots can be visualized on a single gel image and many recent studies involve differential analysis of sets of up to 100 2DE gels, requiring efficient algorithms for this task. Here, we propose a simple \(O(n^2)\) algorithm, exploring the 2DE gel matching problem requirements.

Most of the existing methods for point matching does not take into account the structural information between points to obtain the correspondence. Instead, they usually start by extracting point features (in our case, 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. Existing solutions are based on distances between pair of points (one from each image). For instance, the graduated assignment algorithm, described in (Chui and Rangarajan 2000),
uses the Euclidean distance. Another example is the bipartite graph matching (BGM) combined with shape context (SC), proposed in (Belongie et al. 2002), which is well-known in the literature to be very robust to outliers. Both methods, combined with transform estimation (e.g. thin-plate splines), produce very good matches.

One of the best works, comparing different state-of-the-art methods for the 2DE gel matching problem, is described in (Rogers and Graham 2007). 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, following the same strategy of the ICP algorithm. For the experiments, the authors present detailed evaluation for different distances between points and different point correspondence estimations. They evaluate the Euclidean distance and the SC distance for point matching. Regarding the correspondence estimation, they propose several methods: Closest Point, k-Closest Points, BGM and other graph based ones.

Following this trend, we propose an extension of our previous work described in (Noma et al. 2008), in which it was proposed a quadratic assignment formulation, combining SC as the linear term and structural information as the quadratic term of a cost function, minimized by a correspondence algorithm based on deformed graphs in order to solve the gel matching problem. For complex pairs of gels, we combine an ICP strategy, alternating correspondence and translation estimation, exploring previous knowledge provided by the user, which is fundamental for complex cases, by selecting or validating a couple of corresponding spots. For the experiments, we use simulated pairs to evaluate the performance of our
approach against well-known algorithms involving Euclidean distance, SC, BGM, transform estimation, and ICP, according to different degrees of degradation. Moreover, original pairs, similar to the ones obtained in practice, are also used to assess the results.

3 Framework based on Deformed Graphs

In this section, we present our graph matching framework based on deformed graphs applied to the 2DE gel matching problem.

3.1 Gel Matching

Matching two gel images should be based on invariant features present on them, such as points representing each spot. The best methods for spot matching are based on point matching techniques. Given two sets of 2D points \( x = \{x_1, \ldots, x_N\} \) and \( y = \{y_1, \ldots, y_M\} \), and a distance \( d_{pq} = d(x_p, y_q) \), the idea is to compute the correspondence between points, minimizing a cost function taking into account the distances between corresponding points.

The algorithm proposed in (Almansa et al. 2007) is used for the detection of these points. This algorithm is based on the detection of meaningful spots 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 gel image.
3.2 Graph Matching

We formulate the 2DE gel matching task as a quadratic assignment problem via graph matching. The linear term evaluates the ‘appearance’ while the quadratic term evaluates the ‘structure’. Although graph matching is known to be computationally expensive in general, with sophisticated methods, e.g. (Cour et al. 2007), the proposed framework successfully explores the problem requirements. The general problem of graph matching has a long history and dates back to the 1970’s, e.g. a pioneering work described in (Barrow and Popplestone 1971), which provided proof of concept for the use of relational structures in pictorial object recognition. The authors of (Tsai and Fu 1979) used the attributed relational graphs (ARGs) as a natural data structure to store both kinds of information, appearance and structure. Basically, an ARG is a directed graph in which appearance information is encoded as vertex attributes and structural relations as edge attributes. Given an input and a model ARG, the goal of the graph matching problem is to find a mapping from input to model vertices, evaluating the vertices and edges attributes through a cost function.

Graph matching can be exact or inexact. Exact graph matching is characterized by the edge-preserving restriction in the sense that if two vertices are linked by an edge in the first graph, their corresponding vertices must also be linked by an edge.
in the second graph (Conte et al. 2004). However, in many practical problems, the
edge-preserving condition is too stringent, because the input (or observed) graphs
are subject to deformations such as noise in the acquisition process and intrinsic
variability of the patterns. Instead of prohibiting, the inexact graph matching pe-
nalizes the assignments which are not edge-preserving, increasing the value of the
cost function. Optimal inexact matching algorithms always find a solution that is
the global minimum of the matching cost. If a corresponding exact solution ex-
ists, it will be found by such algorithms. Hence the inexact case can be seen as
a generalization of exact matching algorithms. In both cases (exact and inexact),
the mapping function can represent a homomorphism (injective, many-to-one) or
an isomorphism (bijective, one-to-one). In practice, the isomorphism is too restric-
tive and a weaker form of matching is the subgraph isomorphism, which requires
that an isomorphism holds between one of the two graphs and a vertex-induced
subgraph of the other. An even weaker form of isomorphism is the maximum com-
mon subgraph (MCS), which maps a subgraph of the first graph to an isomorphic
subgraph of the second one. Bunke demonstrated that by a suitable assignment
of costs to edit operations, the MCS problem can be considered a special case of
graph edit distance computation (Bunke 1997). All those matching problems above
are NP-Complete, except for graph isomorphism, for which it has not yet been
demonstrated if it belongs to NP or not. The present paper focus on MCS.

For the graphical representation of each gel image, each point corresponding to
a detected spot is represented by a vertex, as shown in Figure 3(a). ARGs are
used to store both kinds of information, appearance and structure. We focus on
matching two graphs: a model and an input graph. We denote the model graph
by $G_m = (V_m, E_m, \mu, \nu)$, a model vertex by $v_m \in V_m$, its attribute by $\mu(v_m)$, a
model (directed) edge by $e_m \in E_m$ and its attribute by $\nu(e_m)$. Similar notations are
used for the input graph $G_i$. We shall refer all ARGs simply by graphs. Given an input and a model graph representing the two gel images to be matched, the goal is to find a MCS between the two graphs, representing a bijective function (one-to-one correspondence), mapping a subset of input vertices to a subset of model vertices by minimizing the following cost function:

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

The linear term $d_{SC}$ represents the shape context distance and compares pairs of vertices $(v_i, v_m)$ representing corresponding points, directly evaluating the attributes $\mu(v_i)$ and $\mu(v_m)$ (in our case, representing SCs). The quadratic term $d_S$ consists of the structural distance which takes into account geometric penalties. It compares the edge attributes using deformed graphs for efficient evaluation of the structural information encoded by the edges of the graphs. The key idea is to always perform a comparison between graphs having the same topology. Both terms are balanced by a parameter $\alpha$, a real number between 0 and 1.

### 3.3 Shape Context Distance

For the linear term in Equation 1, we use the SC metric inspired by (Rogers and Graham 2007). 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. The normalized histogram at point $p$ is denoted as $h_p(k)$, where the index $k$ identifies the bin. Given this metric we can compute the distance between the SCs of two points $p$ in one image and $q$
in the other image, using the $\chi^2$ distance:

$$d_{SC}(p, q) = d_{\chi^2}(SC(p), SC(q)) = \frac{1}{2} \sum_k \frac{[h_p(k) - h_q(k)]^2}{h_p(k) + h_q(k)}$$  \hspace{1cm} (2)$$

When comparing two SCs, small discrepancies between corresponding points may exist. These discrepancies may have different sources. First, there are genuine differences due to new 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 (Almansa et al. 2007) proposes kernel estimation of the histogram to deal with these discrepancies. This modification of the SC showed robustness and better generalization capabilities.

The spot matching step leads to a matrix $C$ with each entry $C_{pq} = d_{SC}(p, q)$. Therefore, for each spot in one image, we obtain the similarity with each spot in the other one.

### 3.4 Structural Information and Deformed Graphs

Besides SC, structural information is used to improve the classification. We exploit the spatial relations between points representing the detected spots. The key idea is that each directed edge has a corresponding vector as its attribute. Inspired by the work described in (Cesar et al. 2005), the relative positions are evaluated by the following equation, which compares two given vectors, $v_1$ and $v_2$, in terms of the angle between them and their lengths:

$$c_{vec}(\vec{v}_1, \vec{v}_2) = \delta \frac{|\cos \theta - 1|}{2} + (1 - \delta) \frac{||\vec{v}_1| - |\vec{v}_2||}{C_S}$$  \hspace{1cm} (3)$$

in which $\theta$ is the angle between the two vectors and $|\vec{v}|$ denotes the vector modulus (or length).
The first term in Equation 3 represents the ‘angular’ cost, which assigns higher values to ‘opposite’ vectors. The second term represents the ‘modular’ cost, which assigns a value proportional to the difference of the vectors lengths, normalized by a constant $C_S$ to keep the values between 0 and 1. The parameter $\delta$ ranges between 0 and 1.

Given two graphs, one difficulty in the graph matching problem is the fact that in general the two graphs have different sizes, hindering the structural evaluation when comparing edge attributes. Experimental results in the literature show that most of the algorithms produce poor results when matching graphs of different sizes, and this deterioration becomes more evident when this difference in size increases, as described in (Caetano et al. 2006).

For the 2DE gel matching problem, outliers may appear due to differences in protein contents and detection errors. In this case, the input and the model graphs have different number of vertices, making the structural evaluation between input and model graphs costly.

An auxiliary graph, called deformed graph, is used in our approach to overcome this difficulty. Given a pair $(v_i, v_m)$, $v_i \in V_i$, $v_m \in V_m$, in order to evaluate the structural impact when associating an input vertex $v_i$ to a model vertex $v_m$, we first compute the corresponding deformed graph $G_d(v_i, v_m)$, which represents a local deformation in the model graph $G_m$ with respect to edge attributes, caused when ‘simulating’ a change in the coordinates of $v_m$. The deformed graph $G_d(v_i, v_m)$ is computed as follows: it is almost an exact copy of the model $G_m$, with the same number of vertices and edges, and same attributes for the edges, except for those corresponding to the ‘deformed edges’, resulted by replacing the coordinates of $v_m$ by the coordinates of $v_i$ in the corresponding copy of $v_m$ in the deformed graph,
which is denoted by deformed vertex $v_d$. The deformed edges correspond to the edges with an endpoint at $v_d$, as shown in Figure 3(c).

Note that only the edge attributes corresponding to the deformed edges in $G_d(v_i, v_m)$ may differ from the corresponding edge attributes in $G_m$. Therefore, when evaluating edge dissimilarities between the deformed graph $G_d(v_i, v_m)$ and the model $G_m$, only the deformed edges have to be examined, allowing a computationally efficient method.

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

### 3.5 Structural Distance

For the specific problem of gel matching, we used only the structural information in the model graph to compute the correspondences, as shown on the left side of
Figure 3(b). The key idea of deformed graphs is to evaluate the non-linear deformations between the corresponding points. Moreover, ignoring the adjacency information in $E_i$, we avoid topological incompatibilities between input and model graphs. For a given pair $(v_i, v_m)$, firstly we compute its corresponding deformed graph $G_d(v_i, v_m)$. Then, the structural term of Equation 1 can be evaluated by:

$$d_S(G_d(v_i, v_m), G_m) = \frac{1}{|E(v_d)|} \sum_{e_d \in E(v_d)} c_{vec}(\nu(e_d), \nu(e_m))$$

(4)

in which $c_{vec}(\cdot)$ is given by Equation 3, $E(v_d)$ denotes the set of deformed edges with an endpoint at $v_d$, $|S|$ denotes the size of set $S$, $e_m$ is the model edge corresponding to the deformed edge $e_d$, and $\nu(e_d)$ and $\nu(e_m)$ are the vectors corresponding to their edge attributes. Equation 4 is the average cost between deformed edges and their corresponding model edges.

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 side of Figure 3(b). In Figure 3(c), we have an example of large deformation due to a distant $v_i$ from $v_m$.

3.6 Optimization

Both terms in Equation 1 depend only on the current pair being examined and each pair $(v_i, v_m)$ can be evaluated independently by:

$$E(v_i, v_m) = \alpha d_{SC}(v_i, v_m) + (1 - \alpha)d_S(G_d(v_i, v_m), G_m)$$

(5)

and Equation 1 can be rewritten as:

$$E(P) = \sum_{(v_i, v_m) \in P} E(v_i, v_m)$$

(6)
where $P$ is a set of pairs representing an homomorphism between input to model vertices. In order to minimize Equation 6, we use the following greedy strategy. The algorithm receives both input and model graphs, and returns a set $P'$ of pairs $(v_i, v_m)$ representing a MCS between $G_i$ and $G_m$. The possible outliers and spots without correspondence are discarded from the initial solution $P$ 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 $(v'_i, v_m)$ in $P'$. All the remaining pairs $(v_i, v_m), v_i \neq v'_i$ are discarded from $P'$, indicating that they are outliers without correspondence.

**Correspondence Estimation** $(G_i, G_m)$

1. $P \leftarrow \emptyset$
2. for each vertex $v_i \in V_i$
3. do
4. $c_{\text{min}} \leftarrow \infty$; $v_{\text{min}} \leftarrow \text{NULL}$
5. for each vertex $v_m \in V_m$
6. do
7. $c \leftarrow c(v_i, v_m)$
8. if $c < c_{\text{min}}$
9. do
10. $c_{\text{min}} \leftarrow c$; $v_{\text{min}} \leftarrow v_m$
11. $P \leftarrow P \cup \{(v_i, v_{\text{min}})\}$
12. Post-processing of $P$:
   for each $v_m$, keep only the cheapest pair $(v_i, v_m)$ in $P'$.
13. return $P'$

For complex pairs of gels, our approach based on deformed graphs requires a robust estimation to align the input and the model graphs. In the present paper, we explore
ICP, which repeats two steps, correspondence estimation and ‘position updates’, until convergence or a maximum number of iterations is reached. This extension allowed important results using complex pairs of gels and is summarized in the algorithm below:

\[
\text{STRUCTURAL MATCHING}(G_i, G_m)
\]

1. Initialize displacements \(d_0^x\) and \(d_0^y\).
2. \textbf{repeat}
3. Update coordinates in \(G_m\) using \(d_{t-1}^x\) and \(d_{t-1}^y\).
4. \(P^t \leftarrow \text{CORRESPONDENCE ESTIMATION}(G_i, G_m)\).
5. Estimate displacements \(d_t^x\) and \(d_t^y\) from \(P^t\).
6. \textbf{until} convergence or maximum # of iterations
7. \textbf{return} \(P^t\)

In order to align both input and model graphs, we have to estimate a displacement pair \((d_x, d_y)\) ‘which best centralizes’ the model over the input point set, according to the structure evaluated by the deformed graphs. This displacement pair is used to update the model position by adding \((d_x, d_y)\) to each model point coordinate (there is no need to recompute the SC).

As described in (Besl and McKay 1992), the ICP method presents good convergence properties. Nevertheless, it is only guaranteed to converge to a local minimum, and the quality of the result depends on an adequate initialization of the parameters. During the experiments, the median coordinates of each point set were used to calculate the initial values for the displacements \(d_0^x\) and \(d_0^y\): \(d_0^x = \text{median}\{x_i\} - \text{median}\{x_m\}\) and \(d_0^y = \text{median}\{y_i\} - \text{median}\{y_m\}\), in which \(\text{median}\{x_i\}\) is the median coordinate in the \(X\) axis evaluated among all input points, and \(\text{median}\{x_m\}\) is the median coordinate in the \(X\) axis evaluated among all model points. \(\text{Median}\{y_i\}\)
and median \( \{ y_m \} \) are defined analogously.

In order to estimate the displacements \( d^t_x \) and \( d^t_y \) at iteration \( t \), we considered the median displacement on each axis, \( X \) and \( Y \), between the corresponding points in the current estimated mapping \( P^k \), i.e. \( d^t_x = \text{median} \{ x_i - x_m \} \), in which \( x_i \) and \( x_m \) are the coordinates in the \( X \) axis of the corresponding input and model vertices in the mapping (analogously, \( d^t_y = \text{median} \{ y_i - y_m \} \)).

4 Experiments

In this section, the benefits of the proposed algorithm (DG) are compared against some well-known methods in the literature. The graduated assignment (GA) (Chui and Rangarajan 2000) and the bipartite graph matching (BGM) algorithm using shape context (SC) (Belongie et al. 2002), both combined with thin-plate splines (TPS) for the transform estimation.

BGM is the simplest among the tested methods for matching, taking into account only the linear term of Equation 1. (See (Papadimitriou and Steiglitz 1998) for details.) Given a matrix of similarity between spots, \( C_{ij} \), the idea of BGM is to find the optimal assignments which minimizes the cost:

\[
\min_{Q_{ij}} \sum_{ij} C_{ij} Q_{ij}
\]

(7)

where \( Q_{ij} \) is a permutation matrix which encodes the matching. In order to reject the outlier spots, a set of virtual spots with cost \( \varepsilon \) is included expanding the similarity matrix as \( C_{\varepsilon} = [C \ \varepsilon] \). Using ground-truth data, the best results are defined by selecting the parameter \( \varepsilon^* \) which minimizes the number of errors.

In (Rogers and Graham 2007), the authors tested their approach using artificial pairs of point sets. For each artificial pair, they generated two point sets, denoted as
Fig. 4. Experiment 1. Comparison between the proposed algorithm (DG) and BGM (using the hungarian method), using different values for the parameters $\alpha$ and $\epsilon$, respectively, for (a) different percentages of removed points and (b) different degrees of Gaussian noise.

The source was generated using a random uniform distribution, and the target was a distorted version of the source, simulating different degrees of deformations using a spline transformation. In our case, based on real pairs of gel images, we applied the deformations on both images from each pair, using Gaussian noise. In order to simulate the presence of outliers, in (Rogers and Graham 2007), they added and removed points at random from the target set. In our case, we re-
moved points at random from both images from each real pair. For the ground-truth
data, images from (2D gel Datasets) were processed with the method proposed
in (Almansa et al. 2007) and then manually verified. Each gel image is numbered,
e.g. pair “008-009” refers to images “008” and “009” from the dataset.

![Graph a](image1.png) ![Graph b](image2.png)

Fig. 5. Comparison between our approach (DG), BGM, BGM+TPS and GA for (a) different
percentages of removed points and (b) different degrees of Gaussian noise.

The evaluation was divided into three experiments. In the first one, DG is evaluated
against BGM to study the behavior of the parameters $\alpha$ and $\varepsilon$ from DG and BGM,
respectively. The aim of the second experiment is to test the proposed algorithm on
a more challenging situation. We proceeded in a similar way as in the first experi-
ment but using a more complex pair (shown in Figure 6(a)). In this case, we have
a large number of outliers, and strong non-rigid deformations between the two sets
of points, in which all methods fail, while ours demonstrates to be very promising
when using appropriate initial values for $d_0^x$ and $d_0^y$. This is a good feature from the
application point of view, in which it is desired that good results can be achieved
with minimal user interaction. Finally, some results obtained on the original pairs
of gel images from (2D gel Datasets) are shown in Table 1, in order to illustrate the
applicability of our approach to a variety of real situations.
The first experiment follows the same setup described in (Noma et al. 2008). The tests were based on the real pair “008-009” illustrated in Figure 1, which has a greater number of points and outliers, and stronger deformations than the pair “006-007” used in the experiments described in (Noma et al. 2008). In Figure 1, the model is on the left side with $|V_m| = 61$. On the right side, the input has $|V_i| = 56$ and there are 11 input points without correspondences. This experiment was divided into two parts:

**First Part:** Subsets of 10, 20 and 30% were randomly removed from the original sets of points to compare DG against BGM according to increasing number of outliers. This simulates possible errors during detection and natural differences between corresponding spots. For each percentage, 100 trials were considered, and the errors computed for different values of $\alpha$ and $\varepsilon$. In all experiments, the error rate was calculated by taking the number of input vertices with wrong correspondences divided by $|V_i|$.

The parameter $\alpha$ in Equation 5 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 4. The complete behavior for the first part of experiment 1 is presented in Figure 4(a).

**Second part:** In order to increase the nonrigid transformations between points and study the robustness of the algorithm to errors in spot detection, Gaussian noise was added to the points coordinates from both images according to different values of standard deviation $\sigma$, in which higher $\sigma$ corresponds to higher perturbations on the
points coordinates. In this case, both structural and shape information are affected by the Gaussian noise. For each $\sigma$, 100 trials were generated, and the complete behavior for this part of experiment 1 is presented in Figure 4(b).

From these experiments the following conclusion may be drawn: the inclusion of structural information improves the results, in which DG obtains better results with $\alpha$ between 0.2 and 0.8. In this example, for all cases, DG outperformed BGM, demonstrating the importance of the structural information.

Comparing the best results obtained previously, DG with $\alpha = 0.5$ and BGM with $\varepsilon$ between 0.2 and 0.3, against GA and BGM+TPS applied to the same artificial dataset, our approach produced the best results, as shown in Figure 5. In this case, we evaluated the closed-form of TPS transform estimation, represented by the analytical solution for the transformation parameters, using the corresponding pairs of points in the mapping. In Figure 5(a), GA performed poorly, indicating that the Euclidean distance is a poor feature for small point sets with outliers on both sets. On the other hand, SC produced better results. In Figures 5(a) and (b), our results were comparable to BGM+TPS (Belongie et al. 2002), which is known to be very robust to outliers.

4.2 Experiment 2

In this experiment, we illustrate a challenging situation, in which we have a large number of outliers and large non-linear deformations between corresponding points in pair “095-098”, shown in Figure 6(a). On the right side of Figure 6(a), we illustrate the point sets when $\sigma = 5$ which is a more difficult case than the example (Jitter = 4) presented in (Caetano et al. 2006), which corresponds to $\sigma = 4$. Note
Fig. 6. Comparison, using (a) A challenging pair: large number of outliers and strong non-rigid deformations. The model is on the left with $|V_m| = 95$. On the middle, the input has $|V_i| = 119$. There are 57 input points without correspondence. On the right side, an example of Gaussian noise ($\sigma = 5$) on the two point sets which are superimposed, using the initial centralization described in experiment 2 (for curves DG2 in (b)). (b) Comparison between our approach (DG), BGM, BGM+TPS, GA, considering 100 of artificial pairs obtained from (a) using different percentages for points removal and and different standard deviations for Gaussian noise. The curves with and without simulating the prior knowledge given by the user are denoted by DG1 and DG2, respectively.

that in our experiments we handle situations with $\sigma = 10$ and common statistically robust estimators, like RANSAC, are not applicable. The large number of points produces very poor SC information, and even the robust BGM+TPS produced very poor results for this difficult scenario, as illustrated in Figure 6(b). We tested the artificial pairs based on the complex pair “095-098” following the same
strategy of experiment 1, and we observed poor results, as illustrated by curve DG1 in Figure 6(b). In order to test the applicability of the deformed graph to this complex scenario, we simulated the prior knowledge given by the user, initializing $d^0_x$ and $d^0_y$ with the median displacement between the corresponding points from the ground-truth, which improved the results for DG, as illustrated by curve DG2 in Figure 6(b), in which DG outperformed all the other evaluated methods. Thus, more robust results can be achieved when using appropriate $d^0_x$ and $d^0_y$.

Note that, differently from experiment 1, removing larger amounts of points improves the quality of SC and allows BGM to produce better matchings, as shown by the decreasing curve of BGM on the left side of Figure 6(b). This can be explained by the considerable difference between the two original set sizes and by the increasing amount of points without correspondence. On the other hand, BGM+TPS produced poor results which agrees with the fact that the closed-form for the transform estimation is not appropriate for complex cases of gel matching, as described in (Rogers and Graham 2007).

4.3 Experiment 3

Results using original gel images from (2D gel Datasets) are shown in Table 1. For each algorithm, we used the original pairs of gels and computed the corresponding error rates according to the ground-truths. In all examples, our approach (DG) outperformed all the other three methods, BGM, BGM+TPS and GA. The difference in efficiency is more evident in the case of pair “095-098”, in which there is a large number of spots and outliers, and large non-linear deformations between the corresponding points, representing a challenging scenario. Differently from experiment 2, for all those original pairs, including the complex pair “095-
098", we did not use any information from the ground-truth, thus all correspon-
dences were calculated automatically, using $d_x^0 = \text{median}\{x_i\} - \text{median}\{x_m\}$ and
\[d_y^0 = \text{median}\{y_i\} - \text{median}\{y_m\},\]
as described in Section 3.6.

Moreover, we used 21 landmarks given as groundtruth in (2D gel Datasets) for
ten different corresponding images (6...9,12,14...17,19), resulting in 45 possible
combinations of pairs, which were used to illustrate the average performance of
each method. In order to test the presence of outliers, we followed experiments 1
and 2 when removing different percentages of points, generating 100 artificial pairs
for each combination, resulting in 4500 tested pairs (Figure 7). It is important to
note that the original landmarks from dataset (2D gel Datasets) do not correspond
to the maximum concentration of proteins (darkest pixel inside the spot) and do
not include outliers but only the corresponding points, which is not realistic if au-
tomatic methods are used to detect the spots. In Figure 7, for increasing number of
removed spots, the number of outliers increases, degenerating the performance of
all methods while DG2 remains more stable and provides the best results.

Figure 8 illustrates an original pair where our approach (DG) outperforms BGM+TPS
(Belongie et al. 2002), indicating that, for the gel matching problem, the deformed
graphs are more adequate than the closed-form of transform estimations using TPS,
especially in the presence of outliers, typically in real cases.

5 Conclusions

This paper describes a new approach for structural matching of 2D electrophoresis
gels, which is based on two main components: a greedy algorithm for the corre-
spondence estimation and a strategy based on ICP to estimate the coordinates to
Table 1
Comparison between our approach (DG) against Bipartite Graph Matching (BGM, BGM+TPS) and Graduated Assignment (GA), using the original pairs of gels from (2D gel Datasets). The associated parameter values and the corresponding error rates are shown for each method.

<table>
<thead>
<tr>
<th>gels</th>
<th>$\varepsilon$</th>
<th>BGM</th>
<th>$\varepsilon$</th>
<th>BGM+TPS</th>
<th>GA</th>
<th>$\alpha$</th>
<th>DG</th>
</tr>
</thead>
<tbody>
<tr>
<td>006-007</td>
<td>0.3</td>
<td>0.1111</td>
<td>0.2</td>
<td><strong>0.0000</strong></td>
<td>0.1667</td>
<td>0.5</td>
<td><strong>0.0000</strong></td>
</tr>
<tr>
<td>008-009</td>
<td>0.2</td>
<td>0.0492</td>
<td>0.2</td>
<td>0.0179</td>
<td>0.1475</td>
<td>0.5</td>
<td><strong>0.0164</strong></td>
</tr>
<tr>
<td>011-012</td>
<td>0.3</td>
<td>0.0851</td>
<td>0.2</td>
<td>0.1220</td>
<td>0.2927</td>
<td>0.5</td>
<td><strong>0.0244</strong></td>
</tr>
<tr>
<td>014-015</td>
<td>0.2</td>
<td>0.3395</td>
<td>0.2</td>
<td>0.2970</td>
<td>0.1881</td>
<td>0.7</td>
<td><strong>0.1683</strong></td>
</tr>
<tr>
<td>016-017</td>
<td>0.2</td>
<td>0.3684</td>
<td>0.2</td>
<td>0.3878</td>
<td>0.5921</td>
<td>0.3</td>
<td><strong>0.2237</strong></td>
</tr>
<tr>
<td>019-020</td>
<td>0.2</td>
<td>0.0213</td>
<td>0.2</td>
<td>0.0638</td>
<td>0.2340</td>
<td>0.6</td>
<td><strong>0.0145</strong></td>
</tr>
<tr>
<td>028-029</td>
<td>0.2</td>
<td>0.2755</td>
<td>0.2</td>
<td>0.2959</td>
<td>0.3776</td>
<td>0.3</td>
<td><strong>0.1648</strong></td>
</tr>
<tr>
<td>031-032</td>
<td>0.3</td>
<td>0.1539</td>
<td>0.3</td>
<td>0.2692</td>
<td>0.2667</td>
<td>0.1</td>
<td><strong>0.0962</strong></td>
</tr>
<tr>
<td>074-075</td>
<td>0.2</td>
<td>0.1563</td>
<td>0.2</td>
<td>0.2099</td>
<td>0.1875</td>
<td>0.5</td>
<td><strong>0.0617</strong></td>
</tr>
<tr>
<td>095-098</td>
<td>0.2</td>
<td>0.4874</td>
<td>0.2</td>
<td>0.4632</td>
<td>0.6387</td>
<td>0.5</td>
<td><strong>0.1092</strong></td>
</tr>
</tbody>
</table>

“centralize” the model points over the input points. In the correspondence phase, the algorithm minimizes a cost function representing a distance for point comparison, which takes into account structural and shape information.

The proposed algorithm has been tested against well-known methods involving Euclidean distance, SC, BGM, transform estimation, and ICP, using the real pairs available in the public dataset (2D gel Datasets). The test has been divided into
different experiments, in order to demonstrate the proposed method capabilities, showing that the deformed graph is a powerful tool for the specific problem of 2DE gel matching.

The method has been successfully applied to different real image pairs, including complex cases (e.g. pair “095-098”), with larger number of spots and larger deformation between images. In particular, experiment 2 tested artificial pairs based on “095-098”, in which case the results demonstrated that our approach is very effective when using appropriate initial values for the displacements $d_0^x$ and $d_0^y$, thus agreeing with the findings described in (Besl and McKay 1992). The results of our experiments corroborate with the fact that the proposed method is very promising in order to calculate correspondences between complex pairs of gels when using prior knowledge given by the user. Furthermore, the results have also confirmed the need and the robustness of structural information in order to obtain better matchings, as well as the efficiency of the deformed graphs.

Finally, the simplicity of the extended method allows extremely fast implementations. For all tests, our ICP based approach converged in less than 5 iterations (only “095-098” took 10 iterations). Let $n = \max\{|V_l|, |V_m|\}$. Our approach has an upper bound of $O(n^2)$ time complexity, while the asymptotically fastest known BGM
Fig. 8. Pair “031-032”: an example where our approach (DG) outperforms BGM+TPS (Belongie et al. 2002). 1st row: for low values of $\epsilon$, a large number of input vertices are incorrectly discarded as outliers by BGM+TPS (represented by highlighted numbers). 2nd row: for higher values of $\epsilon$, BGM+TPS produces incorrect matches (represented by highlighted lines), indicating that the closed-form of TPS is inadequate. 3rd row: our result.

method is $O(n^{5/2})$ (Papadimitriou and Steiglitz 1998), and GA is $O(n^3)$ (Chui and Rangarajan 2000).

Thus, our method is efficient on both quality and speed, and it can be easily integrated into a graphical interface for user interaction.
Acknowledgements: We are grateful to FAPESP, CNPq, CAPES and PDTSCOP4618. We would like to thank S. Belongie and H. Chui for making their implementations of BGM+TPS and GA, respectively, freely available; and K. A. Nedas for his implementation of the hungarian method for BGM.

References


