Processing of 2D Electrophoresis Gels

Andrés Almansa, Mijail Gerschuni, Alvaro Pardo, Javier Preciozzi

Processing of 2D Electrophoresis Gels*

Andrés Almansa[2], Mijail Gerschuni[1], Alvaro Pardo[1]**, and Javier Preciozzi[1]

[1] DIE, Facultad de Ingeniería y Tecnologías, Universidad Católica del Uruguay
[2] INCO, Facultad de Ingeniería, Universidad de la República.

Abstract. Here we present results on 2D electrophoresis gel image processing using methods that provide a measure of meaningfulness. This work is part of an ongoing project on biomedical image analysis. Biomedical sciences have a long tradition in our country and therefore there is extensive experience in this area. Unfortunately, economical restrictions sometimes do not allow our researchers to have the latest technology. As we will see later, 2D electrophoresis gels are an extremely useful and affordable technology, which can be used in our countries.
From the image processing point of view we believe that there is room for improvements when thinking in the final software application. With this in mind we present an algorithm that covers all the steps for gel image registration. First we present a method for robust and meaningful detection of spots. Then we study two improvements on the computation of the distance between spots using shape contexts. Finally, we present an iterative random sampling process which deals with spot differences between images to give a gel registration. Across all the steps we address the easy interaction with the user based on a measure of meaningfulness of the results.

1 Introduction

This work is part of an ongoing project on biomedical image analysis. One of the goals of this project is to study methods and algorithms for image analysis that provide, together with their results, measures of confidence. That is, we are not only interested in obtaining automatic results with probably some errors; we look for a measure of the confidence on them. Even in the case when the algorithms turn to be semiautomatic, this measure of confidence will be decisive to guide the user towards the issues that need human intervention.

In this paper we address the processing of 2D electrophoresis gel images. 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 isoelectric point and mass. In Fig. 1 we show an example of the kind of images obtained. Each spot in the image represents a protein accumulation and its size depends on

* Supported by Proyecto PDT-S/C/OP/46/18. A. Pardo is on leave from Facultad de Ingeniería, Universidad de la República.
** Corresponding author apardo@ucu.edu.uy.

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, we need to process several images in a single experiment. For this kind of differential analysis we need to register two images finding the spot correspondence. One of the reasons of 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 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 start extracting point features which represent the spots. These point features are then used for point and/or gel matching. In some cases this point features can be used to establish point correspondence before obtaining the complete gel matching [9]. This can be done for example using shape contexts [3]. In [9] the authors review several feature distances and matching procedures, and devise an iterative algorithm to perform the gel registration (based on iterative closest point and graph matching). Some works also include image correlation at some stage to refine the results.

Some of the methods reviewed propose a simple spot detection algorithm which gives a large number of false positives[8]. This is then solved during the matching step with the application of robust methods that discard outliers. We, on the other hand, reject this approach as it produced, in our experiments, too many outliers. We propose a more conservative detection of the spots to increase robustness and reduce computation complexity.

With respect to the use of image correlation we observed that it works for synthetic images, when the same image is compared after deformation, but is does not work in real cases. Real pairs of images not only have strong differences between corresponding spots but also several spots may share the same elliptical appearance.

Finally, although in [9] the authors criticize the use of semiautomatic methods, no matter how precise the methods could be, for this kind of applications some user validation will always be performed. For this reason we present methods which provide information that can be easily used by the operators.

The contributions of this work are the following. First, we present a robust method for spot detection based on level lines that provides a measure of meaningfulness. Second, we include two modifications to the use of shape context which clearly improve the results. Third, we present an iterative sampling process which deals with outliers. After the last two steps we compute a couple of matrices that give the similarity between each pair of spots. These matrices can be also be used by the user as a measure of meaningfulness to validate the results.

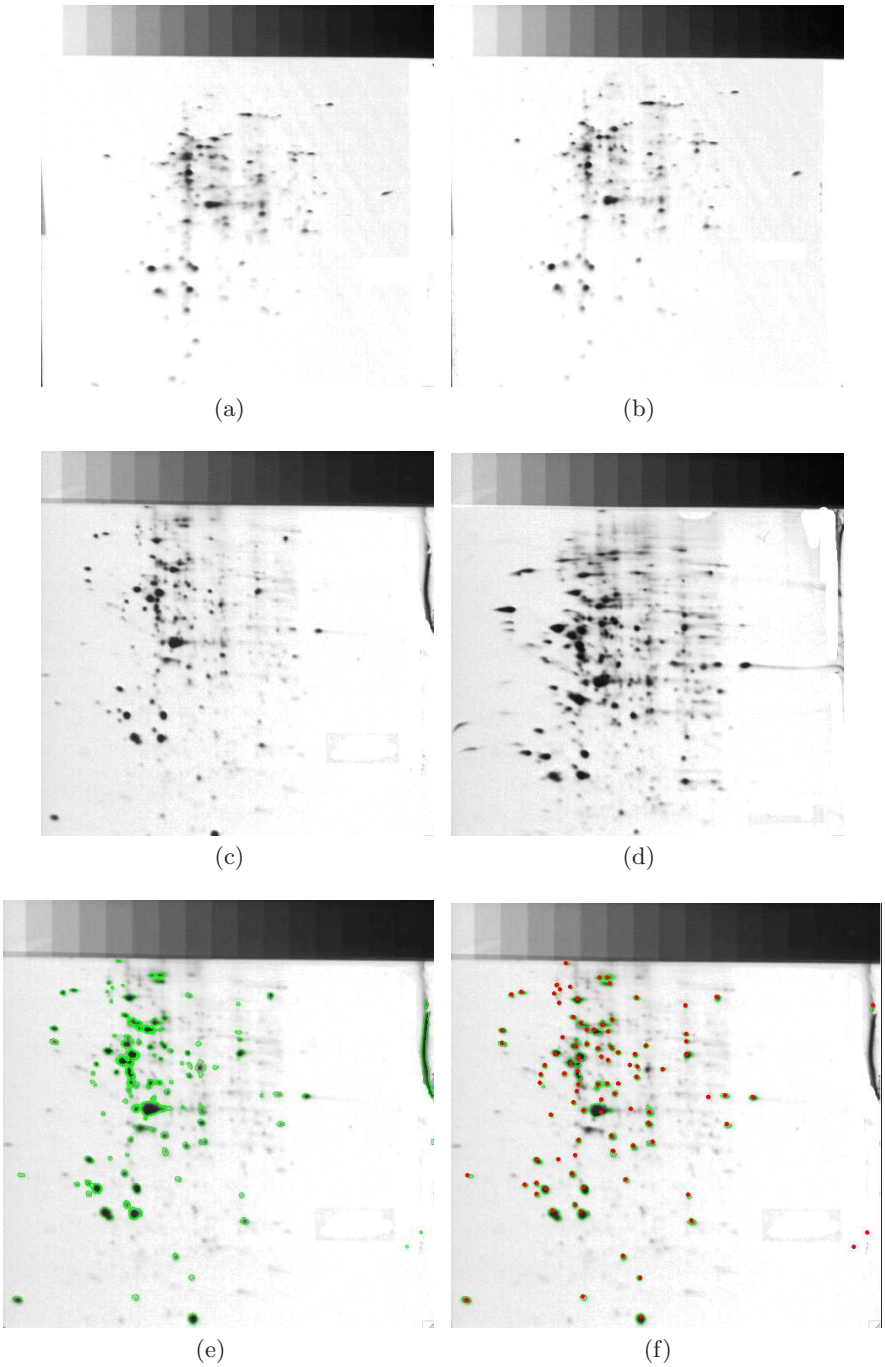

Fig. 1. (a,b) Low complexity pair of images. (c,d) High complexity pair of images. (e) Maximal meaningful boundaries detected of image (d). (e) Resulting shapes obtained applying the isoperimetric relation, over the lower level sets of the image and enters of the spots on the obtained shapes.

2 Spot Detection using Level Lines

The process to establish correspondences between two images must be based on invariant features present on them. Due to the nature of the generation of the gel images explained before, features as the shape or intensity of the spots may vary between experiments, turning them useless. One of the features that remain invariant is the relative distribution of the spots within the image. This is the reason why most of the existing methods are based on point-matching approaches, where the points to be matched are, in general, the center of the spots. Before applying such an algorithm, we must obtain for each image, the set of points to be matched.

The proteins submitted to an electrophoresis process suffer deformations from an initial circular or punctual concentration that turns them into ellipses. Given a set of spots, the points can be obtained as the darkest point of each spot. The problem then is how to obtain the set of valid spots of the image. On this work we propose to detect them using the meaningful boundaries [5] together with several criteria related to the shapes of the spots. We will start in next section with a review of the meaningful boundaries approach.

2.1 Meaningful Boundaries

Meaningful boundaries are based on two concepts: the level lines (connected components of topological boundaries of level sets) of an image [10] and the *a contrario* models [4]. In fact, the meaningful boundaries of an image are obtained by selecting from all of its level lines, those which are meaningful (that is much unexpected) under the *a contrario* model. Lets recall both concepts:

Definition 1 (Level Sets). *Given an image* $u : \Omega(\subset \mathbb{R}^2) \to [a, b](\subset \mathbb{R})$*, we can define for any* $\lambda \in \mathbb{R}$ *the (lower) level sets of* u *as* $\chi_\lambda(u) = \{x \in \Omega, u(x) \leq \lambda\}$.

In the same way we can define the upper level sets. Using both, lower and upper level sets, the *topographic map* of an image is defined as the collection of all of its level lines. Level sets constitute a hierarchical representation of the image which can be encoded in a tree structure, *tree of shapes* [2], to represent the topographic map. Since we are only interested on the detection of dark spots from now on we concentrate only on lower level sets and their corresponding level lines.

***A contrario* models.** Computational Gestalt Theory was first presented by Desolneux, Moisan and Morel [4] as a way to obtain a quantitative theory of the Gestalt laws. Gestalt theory [6] states that visual perception is a grouping process where geometric objects are grouped together by similar features or *gestalts* (color, size, shape, etc). Although the Gestalt theory is consistent from a qualitative point of view, it lacks of a quantitative way to determine when a set of objects have the same gestalt. Computational Gestalt uses the *Helmholtz Principle* to define a quantitative measure of a given gestalt [4].

Helmholtz Principle: The observation of a given configuration of objects in an image is meaningful if the probability of its occurrence by chance is very small.

Therefore we ask the following question: is the observed configuration probable or not in our model?. If not, this proves *a contrario* that this configuration is meaningful. The Helmholtz principle can be formalized by the definition of the *Number of False Alarms* and the definition of an ϵ-*meaningful event*:

Definition 2 (Number of false alarms - NFA and ϵ-meaningful event). *Given an event of the type "a given configuration of objects has a property", the number of false alarms (NFA) is the expectation of the number of occurrences of this event under the uniform random assumption. An event E of the type "a given configuration of objects has a property" is ϵ-meaningful if the NFA is less than ϵ:* $\mathrm{NFA}(E) < \epsilon$.

Since the NFA defined in Def. 2 is in general very difficult to obtain in most practical cases an upper bound definition is used instead. From now on we will use the following re-defintion of NFA.

Definition 3 (Number of false alarms - NFA). *The number of false alarms (NFA) of an event E is defined as:* $\mathrm{NFA}(E) = \mathcal{N} \cdot \mathrm{P}[\mathcal{E} \geq E|H_1]$ *where $\mathcal{N}$ is the number of possible configurations of the event E and H_1 is the background or* a contrario *model.*

Meaningful Level lines. In [5], a definition of meaningful boundary is presented, taking into account the topographic map of the image and the framework reviewed above. Using the norm of the gradient as a measure of the contrast, the authors propose the following ϵ-meaningful boundary definition of a level line C:

$$NFA(C) = N_{ll} P[\min_{x \in C}(|Du(x)|)]^{l/2} < \epsilon$$

where where N_{ll} is the number of level lines in the image, l is the length of C and $P(x)$ is the probability of the contrast x under the *a contrario* model.[3]

A final observation is that on real images, boundaries width are bigger than one pixel, which leads to the detection of several meaningful boundaries for each real one. This problem is solved by applying a maximality criterion over the set of all meaningful boundaries of a monotone section of the tree of shapes.

2.2 Meaningful Spots

Fig. 1(e) shows the result of the meaningful boundaries detection algorithm applied to a real gel image. We can note from the results that most of the spots are detected. Nevertheless, some problems arise. *Shapes that correspond to several spots:* In many cases, several spots are grouped into a bigger one, given as a result, instead of a boundary for each spot, a unique boundary for the whole shape. *Non detected spots:* Some of the spots are not detected as meaningful boundaries. These problems may impact on the point-matching process. In order to solve this problem we must include characteristics of the spots we are looking for. We will address this point in next section.

[3] The power of $l/2$ comes from considering every second pixel's contrsat (or gradient magnitude) on the level line as independent from the others. The *a contrario* model is obtained by using the empirical distribution of the gradient magnitude $|Du|$ as the probability distribution.

2.3 Adding spots characteristics

We already mentioned that we consider only lower level sets in order to detect dark spots. Since most of the spots have an almost elliptical shape we propose to filter the shapes obtained with the meaningful boundaries method using the isoperimetric ratio p^2/a where p is the perimeter and a is the area of the shape. In the case of a circle this relation has a value of 4π.

Although the inclusion of this feature improves the results, some problems are still present. In fact, the problem of detecting the spots that are contained in a bigger one remains unsolved. Let us analyze this case in detail. If several spots are joined on a bigger one two things could happen. If the shape of the grouping has also an elliptic shape it will be detected by the algorithm. On the other hand, if this shape is not elliptical enough to be detected by the algorithm no shape will be detected and some spots will be missed. The first case is shown at Fig. 2, where several spots are grouped on a bigger one with elliptical boundary. In this case, since the criterion to obtain the center of the spot is to keep the darkest pixel, we obtain the center of one of the spots. Nevertheless, the other spot is lost and no information could be obtained from it. In the second case, no inner spot is taken into account, and all this information is lost. In next section we show how to find spot correspondence.

3 Spot Matching

The best methods for spot matching are based on point-matching techniques. In our case we use the metric Shape Context (SC) [3] following [9] 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 two-dimensional histogram in polar coordinates. We will denote the normalized histogram at point i 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:

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

When comparing two SC we must recall that small discrepancies between corresponding points may be encountered. These discrepancies have different sources. First, there are genuine differences due to the appearance or disappearance of spots. Second, the misdetection or errors in spot detection. Third, the gel deformation. To deal with these discrepancies we propose two modifications: a kernel estimation of the histograms and a different metric to compute the distance between SC.

To add robustness and generalization capabilities to the SC we added a kernel estimation of the histogram. This improves the results when small discrepancies are encountered between both images.

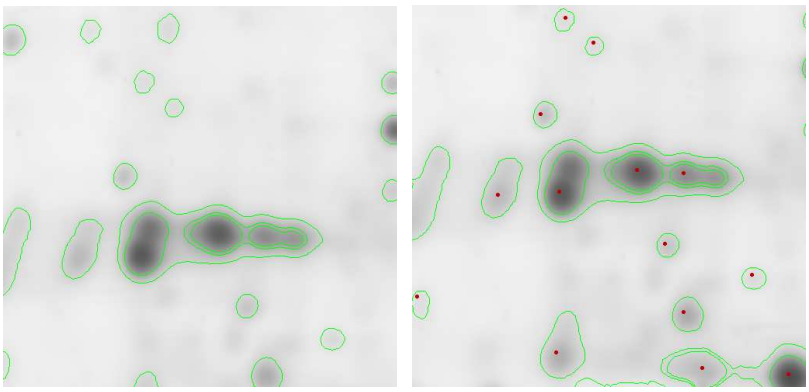

Fig. 2. Extract of a gel image where we can see that two spots are detected on a bigger shape that contains them. We can see also that the center of one of the spots is well detected.

The χ^2 distance it is sensitive to small variations in the histograms. In [7] the authors show how to improve the SC matching using the Earth Movers Distance (EMD). The EMD is a measure of similarity between probability distributions (histograms) that copes with variations. Given two probability distributions S (supply) and D(demand) the EMD distance is the transportation cost from the supply S to the demand D given a cost of transportation between bins: CT_{ij}. Hence the EMD distance between S and D is the minimum effort to transport one probability distribution to the other (For details see [7]). The application of EMD to our problem is relatively straightforward. The probability distributions are the two dimensional SC normalized histograms. The cost of transportation is computed as linear in the radial and circular axis. That is, the distance between bins equals the Euclidean distance between their centers.

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.

To evaluate the performance of both modifications we performed a set of simulations with landmark data obtained from [1]. We used 19 pairs of corresponding images with 399 spots total. In Table 1 we present the results with and without a kernel based estimation of the SC using χ^2 and EMD distances. We show the number of mismatched spots before and after gel registration. The number enclosed in parenthesis corresponds to the results of spot mismatched after gel registration using the method we will propose in section 4. As we can see, the inclusion of the kernel estimation consistently improves the spot matching results. When using χ^2 distance the kernel estimation greatly improves the results. This shows that certain amount of flexibility (generalization) must be allowed in order to capture the intrinsic variations of spots between images. For the same reason, since EMD already allows some discrepancies between SC, the

improvements of the inclusion of the kernel estimation while using EMD are not so impressive. Based on these results we select the number of angle bins, $n_\theta = 24$, the number of radial bins, $n_r = 5$, and outer radius of the neighborhood.

	$n_\theta = 12$	$n_\theta = 16$	$n_\theta = 24$	$n_\theta = 12$	$n_\theta = 16$	$n_\theta = 24$
Kernel EMD	6 (6)	7 (5)	5 (0)	4 (6)	5 (7)	
EMD	9 (6)	6 (14)	5 (11)	4 (9)	5 (5)	
Kernel χ^2	8 (6)	8 (5)	10 (0)	10 (6)	11 (5)	11 (6)
χ^2	24 (6)	21 (14)	25 (11)	23 (9)	19 (7)	29 (15)

Table 1. First three columns: results for outer radius two times the average spot distance. Last three columns: results for outer radius four times the average spot distance. In each cell we show the number of mismatched spots before and after (between parenthesis) gel registration. The last column for EMD is not computed due to its computational cost.

4 Gel Registration

Once we have a list of potential correspondence candidates for each spot we must solve the problem of gel registration. That is, if possible, finding a unique correspondent for each spot. A first approach to obtain such matching could be to find a global set of correspondences which minimize a global merit function. For instance, we could use the method applied in [3]. Given a matrix of similarity between spots, C_{ij}, the idea is to find the optimal assignments to minimize the total cost of matching:

$$\min_{P_{ij}} \sum_{ij} C_{ij} P_{ij} \tag{1}$$

where P_{ij} is a permutation matrix which encodes the matching. Since we may have outlier spots we also include a set of virtual spots with cost ε for rejection purposes. Obviously the problem with the above procedure is that no global coherence is imposed.

As discussed before when pursuing gel registration we have to deal with two types of errors. First we have differences in the spot sets (either genuine or produced during spot). Second, the aforementioned spot differences produce discrepancies in the SC used for spot matching and gel registration. To overcome this problem we developed an iterative random sampling procedure. At each iteration we randomly sample a subset of spots from both images and compute the SC and the corresponding distances between pairs of spots. Along this process, for each pair of corresponding spot we record the smallest distance in a matrix, C_{ij}, and the number of times each pair is matched in a matrix N_{ij}. We use these matrices to compute the gel matching using (1). For each spot in one image we obtain a set of possible corresponding spots. Using C_{ij} and N_{ij} we obtain a measure of confidence of each pairing. This can be used to assist the

user in the rejection of false matches and refinement of the registration solution. No matter how robust the automatic methods could be, there will be always potential mistakes that would need high level information to be resolved. That is why we equipped every step of the method with a measure of confidence to rapidly assist the user.

5 Results and Conclusions

To test the methodology here presented we divided image pairs in three groups of: low, medium and high complexity. In Fig. 3 we show the results. For both pairs we have the ground truth of corresponding points. In the first example of low complexity, in Fig. 3(a), we detected 36 spots in each image and the ground truth contains 29 pairs. The proposed registration process finds 26 correct pairs, 1 erroneous pair and two points with no correspondent. In the second example of medium complexity, in Fig. 3(b), we have 61 spots in one image, 56 in the other one and 45 corresponding pairs in the ground truth. The proposed registration process finds 44 correct pairs and 1 erroneous pair. As we can see the global results are extremely accurate for low and medium complexity pairs. Unfortunately these results are not achieved in high complexity examples. In the example showed in Fig. 3(c) only 25 out of 56 correct pairs are found from 39 spots without correspondent the method correctly finds 22. The lack of global constraints and the complexity of these pairs are not resolved with the proposed method. In this case the strong differences between gels conspire against the results. Although, some of the erroneous correspondents are close to the valid ones, the actual matching is incorrect. This is a clear difference with other works which report the error in pixels. At the end of the day we may have small error but a huge number of incorrect matching to be resolved by the user. Our method intends to overcome this problem and the results for the first two examples seem promising. Below we discuss possible improvements to cope with high complexity pairs.

The spot detection using meaningful boundaries with the addition of the two criteria explained on this article gives good results. Most of the spots present on the images are correctly detected and a very low number of them are missed. Furthermore, we can obtain the meaningfulness for each spot which allows the user to supervise, if needed, the less confident ones. In the future we will address the problem of grouped spots. We expect that the joint inclusion of spots features, instead of in sequential order, will allow us to consider a level line to be meaningful if it is contrasted enough but also has an isoperimetric relation similar to the circle.

The use of a kernel based estimation of the SC histogram and the EMD distance showed to improve the results.

The iterative random sampling process gives good results despite it does not include global coherence. Furthermore, the distance matrix, C_{ij}, and the matrix with the number of matches, N_{ij}, can be used to decide the confidence on the correspondence of pair of spots. We are currently investigating how to include a

more detailed validation of the matching and registration transformation using the theory of Computational Gesltalt. Also, we are exploring the inclusion of constraint on the permutation matrix P_{ij} to reject invalid matchings.

References

1. http://www.lecb.ncifcrf.gov/2dgeldatasets/.
2. C. Ballester, V. Caselles, and P. Monasse. The tree of shapes of an image. Technical report, ENS, 2001.
3. S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(4):509–522, April 2002.
4. A. Desolneux, L. Moisan, and J. Morel. *From Gestalt Theory to Image Analysis. A probabilistic Approach.* Springer "Interdisciplinary Applied Mathematics", 2007.
5. A. Desolneux, L. Moisan, and J.-M. Morel. Edge detection by Helmholtz principle. *Journal of Mathematical Imaging and Vision*, 14(3):271–284, 2001.
6. G. Kanizsa. *La Grammaire du Voir.* arts et sciences. Editions Diderot, 1997.
7. Haibin Ling and Kazunori Okada. An efficient earth mover's distance algorithm for robust histogram comparison. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(5):840–853, May 2007.
8. M. Rogers, J. Graham, and R.P. Tong. 2 dimensional electrophoresis gel registration using point matching and local image-based refinement. In *BMVC04*, volume 2, pages 567–576.
9. M. Rogers and M. Graham. Robust and accurate registration of 2-d electrophoresis gels using point matching. *IEEE Trans. on Image Processing*, 16(3):624–635, March 2007.
10. J. Serra. Image analysis and mathematical morphology. *Academic Press*, 1982.

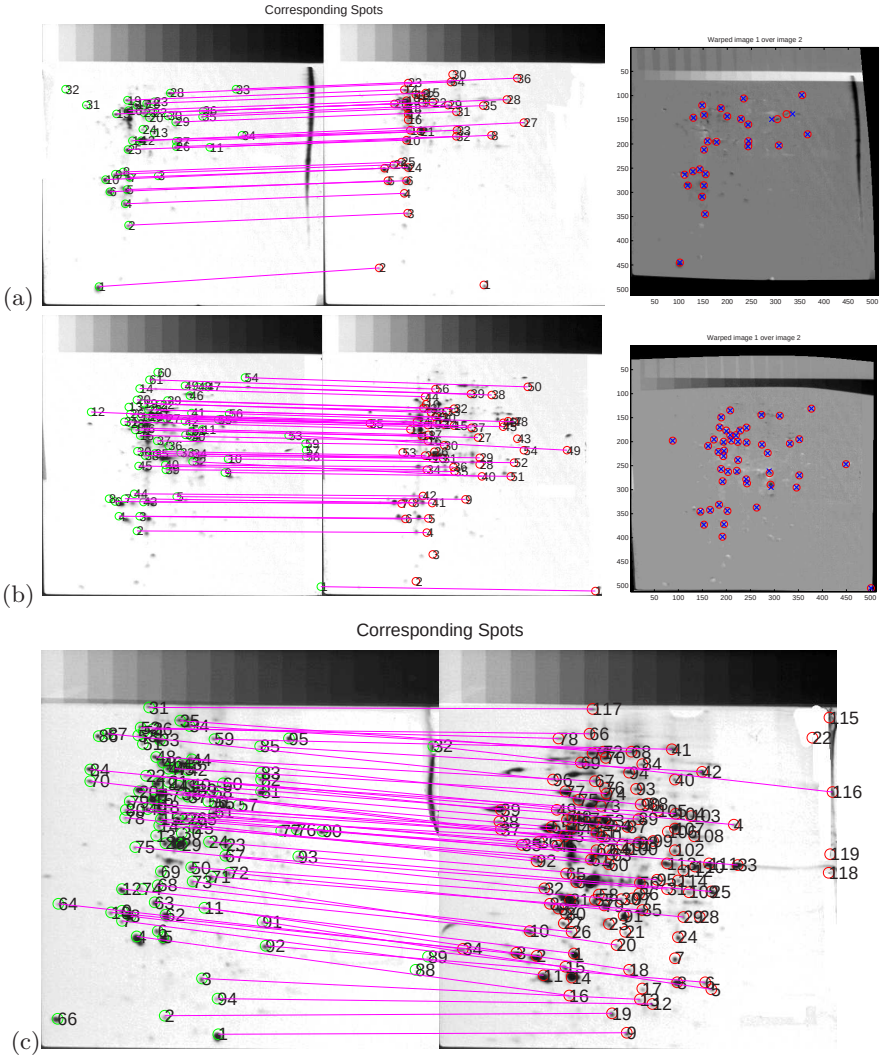

Fig. 3. Results for low (a), medium (b) and high (c) complexity pair of images. (a, b) Left: corresponding pairs. Right registration of both images with thin plate splines. (c) Corresponding pairs.