Improved Resolution of Chromatographic Peak Analysis using Multi-Snapshot Imaging

James R. Hopgood
Institute of Digital Communications, School of Engineering, University of Edinburgh
james.hopgood@ed.ac.uk

***Abstract*—Snapshot imaging has a number of advantages in automated gel electrophoresis compared with the finish-line method in capillary electrophoresis, although at the expense of resolution. This paper presents a novel signal processing algorithm enabling a multi-capture imaging modality which improves resolution. The approach takes multiple snapshots as macromolecules are electrophoresed. Peaks from latter snapshots have higher resolution but poor signal-to-noise ratio (SNR), while peaks from earlier snapshots have lower resolution but better SNR. Signals at different capture-times are related by a scale-in-separation, amplitude scaling, and an arbitrary shift. The multiple captures are realigned and fused together using least-squares estimation and a physically inspired signal model. Since partial waveforms are observed as the chromatic peaks exit the sensor's field-of-view, this is accounted for in the realignment algorithm. The proposed technique yields improved resolution, improved fragment concentration and size estimates, and allows the removal of static background noise.**

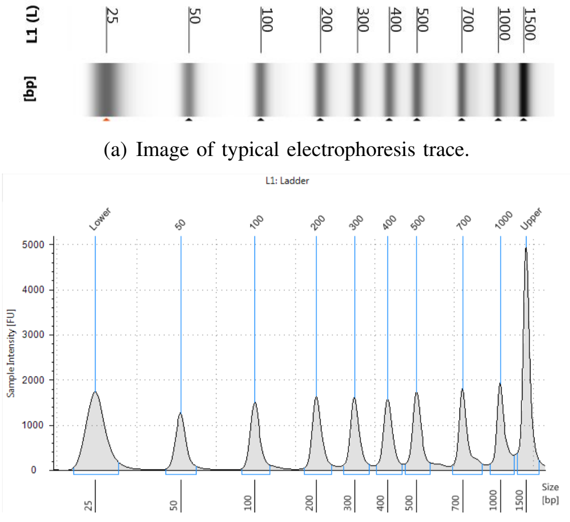
***Index Terms*—Chromatography, snapshot imaging, finish-line method, least-squares estimation, signal modelling.**

I. Introduction

Electrophoresis is a fundamental and ubiquitous technique for separating individual macromolecules in biological samples, such as DNA, RNA, and proteins. It is used in forensics, molecular biology, genetics, microbiology and biochemistry. While next generation sequencing (NGS) facilitates the precise order of nucleotides within a DNA sequence, gel electrophoresis has a number of important applications: in particular, as an economically viable quality control stage for NGS; for the analysis of proteins; or genetic fingerprinting, among others.

There are two common detection methods for measuring the separation of macromolecules in electrophoresis, finish-line and snapshot imaging [1]. In the finish-line method, molecules are detected after electrophoresis for a constant distance using a single point detector at some fixed point in space. In snapshot imaging, molecules are electrophoresed for a constant time, calibrated such that the sample has expanded to fit the imaging sensor's entire field of view, thereby maximising the use of the sensor's spatial resolution (see Fig. 1); the entire electrophoresis trace is therefore imaged after some fixed duration. In both cases, the calibration is usually performed using a ladder sequence for benchmarking. Finish-line methods generally have the best resolution as a result of molecules running to a constant distance [2].

Automated electrophoresis processes are ubiquitous in the genomics lab. Most commercial devices either use capillary electrophoresis (CEP) with finish-line measurement, or gel electrophoresis (GEP) using snapshot imaging. While CEP methods have excellent resolution, automated CEP technologies tend to be slow, expensive, and have weaknesses compared to snapshot-based GEP methods. For example: commercial CEP products usually run samples sequentially, rather than in parallel, thus taking more time to analyse multiple samples; channels reused for several samples are prone to *carry-over* or potential contamination between samples; and the use of a single-detector means a molecule is observed only once, so high dynamic range techniques cannot be used. Snapshot GEP does not suffer these restrictions. Therefore, in order to benefit from the advantages of automated snapshot GEP, it is desirable to develop techniques that build on snapshot imaging and provide the resolution of finish-line imaging methods.



(a) Image of typical electrophoresis trace.

(b) A 1D electrophoresis signal.

Fig. 1. Snapshot of an electrophoresis trace of a standard DNA ladder.

This paper develops a signal processing framework for a multi-capture technique. In the proposed approach, information from multiple images is fused together to provide a single high-resolution image at a standard electrophoresis (EP) time. This technique is timely due to improvements in imaging technology, hardware-based processing capability, and advances in automation processes on commercial products.

The proposed multi-capture or multi-snapshot imaging (MSI) method provides a number of benefits to the analysis

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

of the macromolecules over single-snapshot imaging (SSI) and indeed the finish-line method. These include improved resolution, improved estimation of fragment concentration and size, and removal of static background noise, or observation noise due to discarded fluorescence. While many algorithms have looked at a variety of aspects for improving the analysis of gel electrophoresis [3]–[5], the technique of multi-capture has neither previously been published, nor previously implemented in a commercial product.

II. Motivation for Multi-Capture Imaging

An example of snapshot imaging in GEP is shown in Fig. 1(a). The central image rows are used to create an 1D signal, as in Fig. 1(b), where the x-axis denotes either uncalibrated molecular weight, base-pair fragment length, or more generally *separation*, with smaller particles on the left, larger ones on the right.

A. Separation Resolution Definition

To see how MSI improves resolution, consider the simplest model for describing the evolution of the concentration field. This model, used in both the analysis of finish-line and snapshot detection methods, considers the evolution of a unit mass injected as a delta function at the origin of an electrophoresis system [1]. The concentration field, $c(x,\, t)$, resulting from this injection is, at time t and separation x, the solution to the averaged convection-diffusion equation for large-times,

$$\frac{\partial c}{\partial t} + \bar{U}\frac{\partial c}{\partial x} = \bar{D}\frac{\partial^2 c}{\partial x^2} \tag{1}$$

where $\bar{U}$ is the mean velocity vector of the injected molecules, and $\bar{D}$ the dispersion coefficient. The solution in this idealised case is a Gaussian pulse [1]:

$$c(x,\, t) = \frac{1}{\sqrt{2\pi\sigma_t^2}} \exp\left[-\frac{(x-\mu_t)^2}{2\sigma_t^2}\right] \tag{2}$$

where the mean position $\mu_t = \bar{U}t$, and $\sigma_t^2 = 2\bar{D}t$ is the *variance* of the peak. It is crucial to note that the position of the peak increases linearly with time, while the peak's full-width at half maximum (FWHM) increases sub-linearly with time, namely $W_t = 2\sqrt{2\ln 2}\,\sigma_t = 4\sqrt{(\ln 2)\,\bar{D}t}$ varies with the square root of time. This means the peaks separate more quickly than they broaden.

The separation resolution between two peaks, $i \in \{1, 2\}$, with means $\mu_{t,i}$ and variances $\sigma_{t,i}^2$ is defined as [1], [6]:

$$R_{s,t} \triangleq \frac{|\mu_{t,1} - \mu_{t,2}|}{2\left(\sigma_{t,1} + \sigma_{t,2}\right)} \tag{3}$$

B. Realignment Improves Resolution

In order to obtain an improvement in resolution with snapshot imaging to match the finish-line method, it is necessary to run the electrophoresis for longer, and then map the new trace back to the standard EP time. To see this, consider the concentration field at times $t = t_1$ and $t = t_2$, as shown in the upper two graphs in Fig. 2. This figure is representative of two simple symmetric clean peaks seen in electrophoresis traces. At $t = t_k$, for $k = \{1, 2\}$, the two peaks have mean position $\mu_{t_k,i} = \bar{U}_i t_k$ and peak deviation of $\sigma_{t_k,i} = 2\bar{D}_i t_k$. Suppose the waveform at time $t = t_2$ is rescaled in the separation axis by a factor of $\Delta_{12} = \frac{t_1}{t_2}$, yielding the bottom graph in Fig. 2. The mean positions of the rescaled waveform are shifted to:

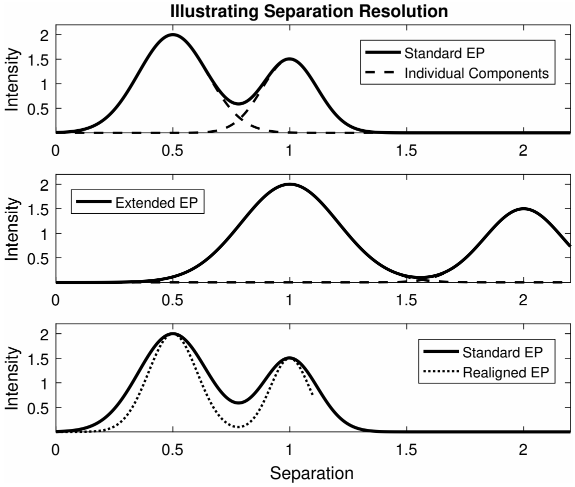


Fig. 2. Illustration of resolution change as electrophoresis continues. In this figure, the electrophoresis process moves to the right, in contrast to Fig. 1.

$$\hat{\mu}_{t_2,i} = \Delta_{12}\,\mu_{t_2,i} = \frac{t_1}{t_2}\bar{U}_i t_2 = \bar{U}_i t_1 = \mu_{t_1,i} \tag{4}$$

meaning the rescaled peak positions match the original peak positions. Moreover, the rescaled peak widths are:

$$\hat{\sigma}_{t_2,i} = \Delta_{12}\,\sigma_{t_2,i} = \frac{t_1}{t_2}\sqrt{2\bar{D}_i t_2} = \sqrt{\frac{t_1}{t_2}}\sigma_{t_1,i} \tag{5}$$

Therefore, the resolution of the rescaled waveform is:

$$\hat{R}_{s,t_2} \triangleq \frac{|\hat{\mu}_{t_2,1} - \hat{\mu}_{t_2,2}|}{2\left(\hat{\sigma}_{t_2,1} + \hat{\sigma}_{t_2,2}\right)} = \sqrt{\frac{t_2}{t_1}} R_{s,t_1} \tag{6}$$

Since $t_2 > t_1$, then the resolution $\hat{R}_{s,t_2} > R_{s,t_1}$ has improved, as shown by the narrower peaks in the lower figure of Fig. 2. These results can easily be extended to the case where an impulse is injected at time $t = t_0$ at separation $x = x_0$, with the times being replaced by $t_i - t_0$. In practice, due to a number of effects, the peak shapes are not Gaussian as given by equation (2), although a number of other shapes such as Voigt, pseudo-Voigt, and asymmetric variants are frequently used [5]. Nevertheless, the resolution improvement from this realignment process applies irrespective of the peak shape.

C. Fusing Realigned Traces

In Fig. 2, the center graph shows the concentration field when electrophoresis is continued after the standard electrophoresis time. It is seen that the right-most peak (with the smalled size) is exiting the field of view and therefore, when realigned, only partially overlaps the standard EP trace as shown in the bottom graph of Fig. 2. While the realigned waveform has improved resolution, there is missing information for separations $1.1 < x \leq 2$. Therefore, to benefit from

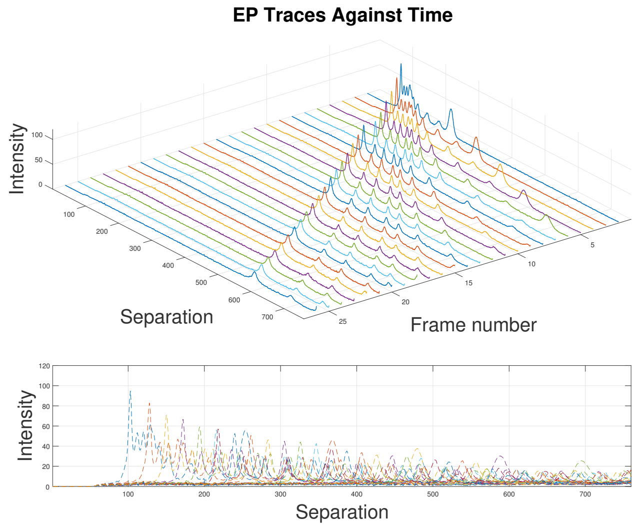


Fig. 3. Original electrophoresis snapshots; sampling period of 60 seconds.

the improvement in resolution by running electrophoresis for longer, fusion of the different realigned pulses is required.

A simple approach to fusing the realigned traces is to take, at a given separation, x, the realigned signal corresponding to the most recent capture. However, as the EP time increases, the signal-to-noise ratio (SNR) decreases due to the reduction in peak amplitude. Therefore, it would be preferable to use all the information available. Moreover, such brute-force stitching leads to discontinuities in the fused waveform. Other approaches such as averaging the waveforms leads to band broadening and loss of resolution. A solution for this realignment is discussed in section §III-C.

III. Proposed Algorithm

The realignment mechanism needed in MSI is equivalent to estimating the scaling factor and shift, and is a classic estimation problem. Carlson [7] produced an algorithm for problems involving Doppler shifts, although it assumed the entire waveform is present in both measurements. In MSI, the situation is complicated by the fact the signal leaves the observation window, and therefore matching is required over partial signals. As a result, direct application of cross-correlation does not give accurate realignment. This section derives the full algorithm, starting by developing the appropriate expansion model used to derive the signal model in section §III-B.

A. Linear Chromatography Expansion Model

If the electrophoresis process is continued past the standard EP time, the macromolecules continue to separate, as shown in the example traces in Fig. 3. If the electric field is constant, and the behaviour of the gel remains invariant, the convection-diffusion equation (1) indicates a linear expansion model, since the mean-position of the injected molecules increases linearly with time, as shown in equation (2) of section §II-A. This expansion is shown diagrammatically in Fig. 4.

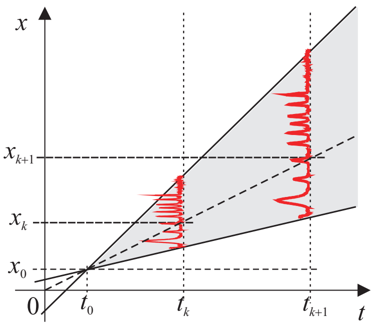


Fig. 4. Linear GEP expansion model. This is a rotated schematic of Fig. 3.

Suppose a sample is injected such that the origin of the GEP expansion begins at $(x_0,\, t_0)$. A macromolecule at *any point* $(x_k,\, t_k)$, denoted by $c(x_k,\, t_k)$, will later appear at position $c(x_{k+1},\, t_{k+1})$ through a simple geometric mapping depending on the expansion model. In the linear case, from similar triangles (see Fig. 4), then $c(x_k,\, t_k) \propto c(x_{k+1},\, t_{k+1})$, and the position of a molecule at t_k and t_{k+1} are related through:

$$\frac{x_{k+1} - x_0}{t_{k+1} - t_0} = \frac{x_k - x_0}{t_k - t_0} \tag{7}$$

or, equivalently, writing $\Delta_k = \frac{t_{k+1} - t_0}{t_k - t_0}$ and $\Sigma_k = \frac{t_{k+1} - t_k}{t_k - t_0}$,

$$x_{k+1} = \Delta_k\, x_k + \Sigma_k \tag{8}$$

This means the signal value from the electrophoresis trace at a given position x at times t_k and t_{k+1} are related by:

$$c(x,\, t_{k+1}) = \alpha_k\, c\left(\frac{x - \Sigma_k}{\Delta_k},\, t_k\right) \tag{9}$$

where α_k is a scaling coefficient. Note that $\Sigma_k = (1 - \Delta_k)\, x_0$. Thus, if $x_0 \geq 0$, and the scaling factor $\Delta \geq 1$, then the shift $\Sigma_k \leq 0$, while $x_0 < 0 \Rightarrow \Sigma_k > x_0$.

Given the initial injection point, $(x_0,\, t_0)$, the scale-in-separation can be evaluated from one frame to the next, $\Delta_k = \Delta_k(t_0,\, t_k,\, t_{k+1})$, as can the amplitude scaling α_k, and the shift $\Sigma_k = \Sigma_k(x_0,\, t_0,\, t_k,\, t_{k+1})$. However, it is not always possible to accurately determine $(x_0,\, t_0)$, and therefore the position of the injection point is assumed unknown.

Moreover, while $(x_0,\, t_0)$ can be estimated from a number of multiple-snapshots assuming linear expansion, in practice, the expansion model is nonlinear due to Joule heating, ionization of the gel, deterioration of the fluorescent dyes, and other effects. Therefore, the scale-in-separation, amplitude-scaling, and shift should be estimated on a per-snapshot basis. These parameters can estimated efficiently using the ubiquitous least-squares estimate (LSE) approach, as shown in section §III-B.

B. LSE Parameter Estimation

Consider realigning snapshots (frames) at times $t = t_P$ and $t = t_R$, the so-called *projected* and *reference* frames; the former will be realigned to the latter reference frame. First assume $t_R > t_P$, and define $f_k[n] = c(n\delta x,\, t_k)$ to indicate a spatially quantised version of the concentration field at separations $x = n\delta x$ for $n \in \{0, \ldots, N-1\}$, with $\delta x = \frac{1}{N} x_{\max}$, where $x_{\max}$ is the maximum separation.

a) Linear Model: In the simplest model, the reference signal is modelled as a linear multiple of the scaled-and shifted projected frame with a modelling error $e[n]$:

$$f_R[n] = \alpha\, f_P[n;\, \Delta,\, \Sigma] + e[n] \tag{10}$$

where $f_P[n;\, \Delta,\, \Sigma] = f_P\left[\frac{n\delta x - \Sigma}{\Delta}\right]$, as given by (9). In order to estimate Δ, Σ, and α, it makes intuitive sense to optimise their values to minimise the average square modelling error. Define the average error as:[1]

$$E_T\left(\boldsymbol{\theta}\right) = \frac{1}{N}\sum_{n=0}^{N-1} e^2[n] = \frac{1}{N}\sum_{n=0}^{N-1}\left(f_R[n] - \alpha\, f_P[n;\, \Delta,\, \Sigma]\right)^2 \tag{11}$$

The objective is to find the set of parameters $\boldsymbol{\theta} = \{\Delta,\, \Sigma,\, \alpha\}$ that minimises $E_T\left(\boldsymbol{\theta}\right)$; in other-words:

$$\boldsymbol{\theta}_{\text{opt}} = \arg\max_{\theta} E_T\left(\boldsymbol{\theta}\right) \tag{12}$$

The gain, α, can be found analytically given Δ and Σ:

$$\alpha = \frac{\sum_{n=0}^{N-1} f_R[n]\, f_P[n;\, \Delta,\, \Sigma]}{\sum_{n=0}^{N-1} f_P^2[n;\, \Delta,\, \Sigma]} \tag{13}$$

The total error in equation (11) can thus be written as:

$$\epsilon_T\left(\boldsymbol{\theta}\right) = \sum_{n=0}^{N-1} f_R^2[n] - \frac{\left\{\sum_{n=0}^{N-1} f_R[n]\, f_P[n;\, \Delta,\, \Sigma]\right\}^2}{\sum_{n=0}^{N-1} f_P^2[n;\, \Delta,\, \Sigma]} \tag{14}$$

where $\epsilon_T\left(\boldsymbol{\theta}\right) = N E_T\left(\boldsymbol{\theta}\right)$. The total error in (14) can be minimised with respect to $\{\Delta,\, \Sigma\}$: a nonlinear LSE problem.

b) Affine Model: The model in (10) doesn't model the baseline present throughout the entire electrophoresis process, but can be modified to include the baseline as follows:

$$f_R[n] = \alpha_0 f_P[n;\, \Delta,\, \Sigma] + \sum_{q=1}^{Q} \alpha_q\, g_q[n] + e[n] \tag{15}$$

where $\{\alpha_q\}_1^Q$ are unknown coefficients, and $g_q[n]$ are Q known basis functions representing the general shape of the baseline. Define the gains by $\boldsymbol{\alpha} = \left[\alpha_0,\, \alpha_1,\, \cdots,\, \alpha_Q\right]^T$, and the augmented *projected signal*:

$$\mathbf{f}_S[n;\, \Delta,\, \Sigma] = \left[f_P[n;\, \Delta,\, \Sigma] \quad g_1[n] \quad \cdots \quad g_Q[n]\right]^T \tag{16}$$

such that $f_R[n] = \boldsymbol{\alpha}^T \mathbf{f}_S[n;\, \Delta,\, \Sigma] + e[n]$. Defining:

$$E_T\left(\boldsymbol{\theta}\right) = \frac{1}{N}\sum_{n=0}^{N-1}\left(f_R[n] - \boldsymbol{\alpha}^T\, \mathbf{f}_S[n;\, \Delta,\, \Sigma]\right)^2 \tag{17}$$

then as in equation (12), the objective is to find the set of parameters $\boldsymbol{\theta} = \{\boldsymbol{\alpha},\, \Delta,\, \Sigma\}$ that minimises $E_T\left(\boldsymbol{\theta}\right)$.

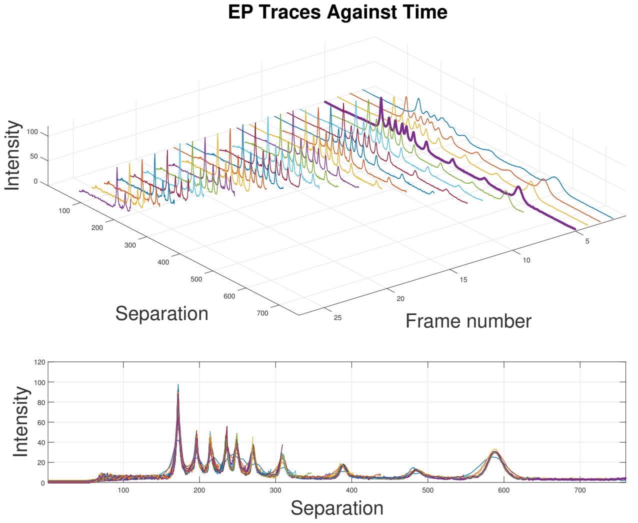


Fig. 5. Realigned electrophoresis traces using LSE and the affine model.

The optimal gain vector $\boldsymbol{\alpha}$ satisfies:

$$\underbrace{\left[\sum_{n=0}^{N-1} \mathbf{f}_S[n;\, \Delta,\, \Sigma]\ \mathbf{f}_S^T[n;\, \Delta,\, \Sigma]\right]}_{\mathbf{R}\,[\Delta,\,\Sigma]} \boldsymbol{\alpha} = \underbrace{\sum_{n=0}^{N-1} \mathbf{f}_S[n;\, \Delta,\, \Sigma]\, f_R[n]}_{\mathbf{r}\,[\Delta,\,\Sigma]} \tag{18}$$

and the total error in equation (17) is:

$$\epsilon_T\left(\boldsymbol{\theta}\right) = \sum_{n=0}^{N-1} f_R^2[n] - \mathbf{r}^T\left[\Delta,\, \Sigma\right]\mathbf{R}^{-1}\left[\Delta,\, \Sigma\right]\,\mathbf{r}\left[\Delta,\, \Sigma\right] \tag{19}$$

where $\epsilon_T\left(\boldsymbol{\theta}\right) = N E_T\left(\boldsymbol{\theta}\right)$. As in equation (14), the total error in (19) can be minimised with respect to $\{\Delta,\, \Sigma\}$ using gradient descent, or a grid search.

C. Fusing Realigned Traces

The realigned traces must be fused. A principled approach for this fusion is to consider how the electrophoresis trace is predicted to change over time using the averaged convection-diffusion equation (1). After realignment, the concentration field is given by (2) with μ_t and σ_t^2 replaced by (4) and (5). This model, however, doesn't take into account loss of fluorescent dye, nor the asymmetric peak shapes that result in practice. Instead, a fusion method found to give good results is one which models the *local* concentration fields across the most recent captures as a Gaussian with an offset:

$$\hat{\mathcal{C}}(x,\, t) = A_{x,0} + \frac{A_{x,1}}{\sqrt{2\pi\sigma_x^2}} \exp\left[-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right] \tag{20}$$

The parameters $\{A_{x,0},\, A_{x,1},\, \mu_x,\, \sigma_x^2\}$ are estimated at each separation x by fitting, in a least squares sense, $\hat{\mathcal{C}}(x,\, t)$ to the observed measurements across a window of separation values $\{x - N_w,\, \cdots, x-1,\, x,\, x+1, \cdots,\, x+N_w\}$ and across the most recent N_T snapshots, subject to boundary conditions. The concentration is then evaluated using equation (20) and the estimated parameters at the actual separation x.

[1] Since $t_R > t_P$, then $\Delta \geq 1$ and the domain of $f_R[n]$ covers the domain of $f_P[n]$, and hence the summation over the N discrete-separation values.

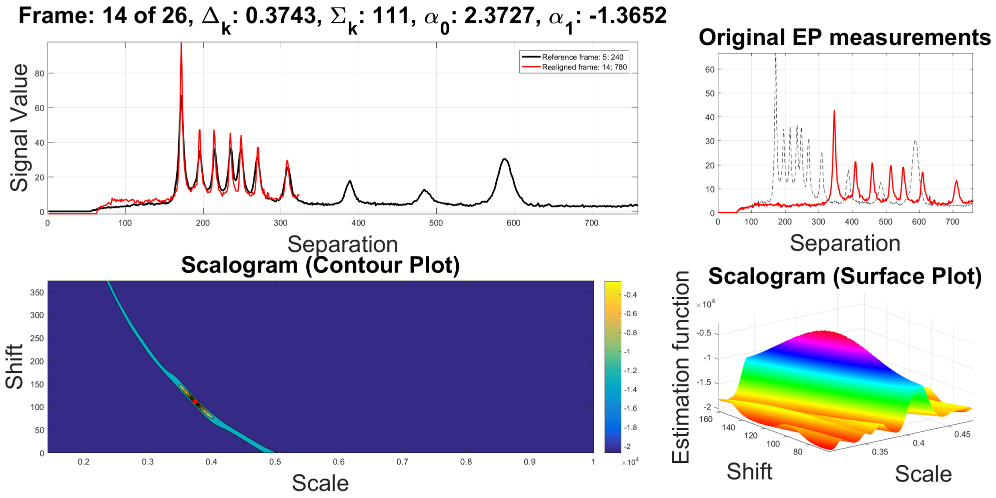


Fig. 6. Alignment of two frames, and the negative error surface of equation (19) as a function of shift and scale. Here, the negative error-surface is called the scalogram. A peak in the scalogram indicates the best fit realignment parameters.

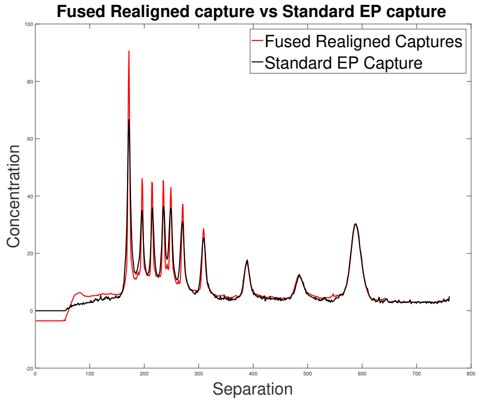


Fig. 7. Comparison of the reference electrophoresis trace and the fused realigned traces using the method in section §III-C.

IV. Results

To demonstrate the realignment process, a typical DNA sample is analysed. In this experimental setup, snapshots where taken at multiples of 30 seconds for 750 seconds. The standard EP time is 180 seconds, corresponding to the fifth frame. The resulting multiple snapshots are shown in Fig. 3. In the affine model described in section §III-A, $Q = 2$ indicating that a linear baseline model $(\alpha_1 + \alpha\, n)$. The scaling Δ, shift Σ, gain α_0, and baseline coefficients $\{\alpha_q\}_1^Q$ are estimated by minimising (19) with respect to the fifth frame, as described in section §III-B. An example of the resulting error surface from equation (19) as a function of shift and scale is shown in Fig. 6 when the 14th frame is realigned to the fifth frame. The realignment of all the frames is shown in Fig. 5. Fig. 7 shows the fused trace, as compared with the standard electrophoresis trace. The realigned-and-fused trace shows improved SNR and improved resolution, as indicated by the sharper peaks.

V. Conclusions

This paper presents a novel multi-capture snapshot imaging technique for GEP using estimation theory to realign and fuse multiple waveforms. The fused waveform demonstrates both improvements in resolution and sensitivity due to increased SNR. A further advantage of this technique is that static noise will be diminished in the fusion. Improvement in resolution can be quantified by deconvolving the electrophoresis traces, so that individual peak separation and widths are known, followed by using (3). This quantification is reported elsewhere.

Acknowledgment

This work has been funded by the Royal Academy of Engineering through its Industrial Secondment Scheme.

References

[1] K. D. Dorfman, S. B. King, D. W. Olson, J. D. P. Thomas, and D. R. Tree, "Beyond gel electrophoresis: Microfluidic separations, fluorescence burst analysis, and dna stretching," *Chemical Reviews*, vol. 113, no. 4, pp. 2584–2667, Nov. 2013.

[2] J. C. Sutherland, D. J. Fisk, D. C. Monteleone, and J. G. Trunk, "A comparison of electrophoretic resolution for snapshot and finish-line imaging," *Analytical Biochemistry*, vol. 239, no. 2, pp. 136–144, Aug. 1996.

[3] J. W. Yoon, S. J. Godsill, C. Kang, and T.-S. Kim, *Proceedings of Bioinformatics Research and Development: First International Conference*. Germany: Springer Berlin Heidelberg, Mar. 2007, ch. Bayesian Inference for 2D Gel Electrophoresis Image Analysis, pp. 343–356.

[4] N. Kaabouch, R. Schultz, and B. Milavetz, "An analysis system for DNA gel electrophoresis images based on automatic thresholding an enhancement," in *Electro/Information Technology, 2007 IEEE International Conference on*, May 2007, pp. 26–31.

[5] V. B. D. Marco and G. G. Bombi, "Mathematical functions for the representation of chromatographic peaks," *Journal of Chromatography A*, vol. 931, no. 1-2, pp. 1–30, Oct. 2001.

[6] A. Felinger, *Data Analysis and Signal Processing in Chromatography*. Elsevier Science, May 1998.

[7] J. E. Carlson and F. F. Sjoberg, "Simultaneous maximum likelihood estimation of time delay and time scaling," in *Proceedings of the 6th Nordic Signal Processing Symposium (NORSIG)*, Jun. 2004, pp. 260–263.