COMP0118: Respiratory Motion Modelling

Author: Ciaran Coleman (909473)

April 21, 2021

1 Introduction

Respiratory motion proves problematic for a range of medical procedures that rely on accurate image acquisitions to diagnose, plan or assist with an intervention. The motion deforms surrounding organs and tissues, leading to occlusion (among other motion artefacts), as well as increased uncertainty in tracking anatomical locations. For medical interventions, MR imaging is not yet able to provide high-quality, accurate images in real-time that captures the internal motion caused by respiration [10]. Surrogate-driven respiratory motion models potentially solve this by first building a correspondence model using synchronised internal imaging and easy-to-measure surrogate signals (e.g. markers on the skin surface). Once built, only the surrogate signals need to be tracked to estimate the internal motion.

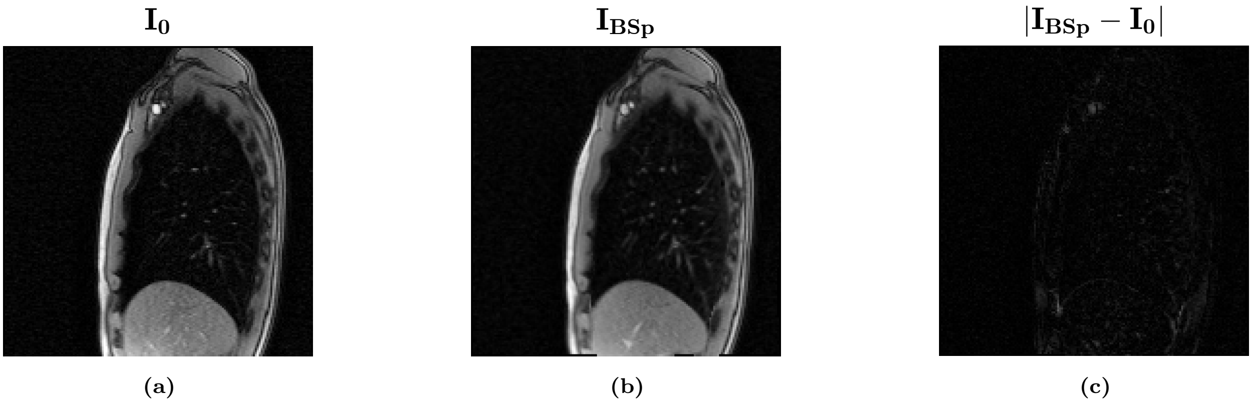
In this report, numerous correspondence models are fit to 2D saggital MR images to estimate the internal motion of a subject thorax and abdomen. Both skin surface and diaphragm surrogates (as well as time derivatives) are considered. The models are quantitatively evaluated by estimating the internal motion using the models and comparing with that from B-spline image registration and trends in performance over time investigated.

2 Materials and Methods

2.1 Data

The dataset consists of 1500 sagittal MR images from a single subject, acquired at around 3fps and spanning $\approx 500\,\mathrm{s}$ of time in total. An example image is shown in Figure 1a. B-spline image registrations were provided.

2.1.1 Registration Results



(a) **(b)** **(c)**

Figure 1: Visual assessment of B-spline registrations: (a) Target image at end-inhalation, (b) Source image deformed using B-spline transformation, (c) Absolute intensity difference between target and estimated. Intensity scale is the same on all three images to emphasise similarity in (c).

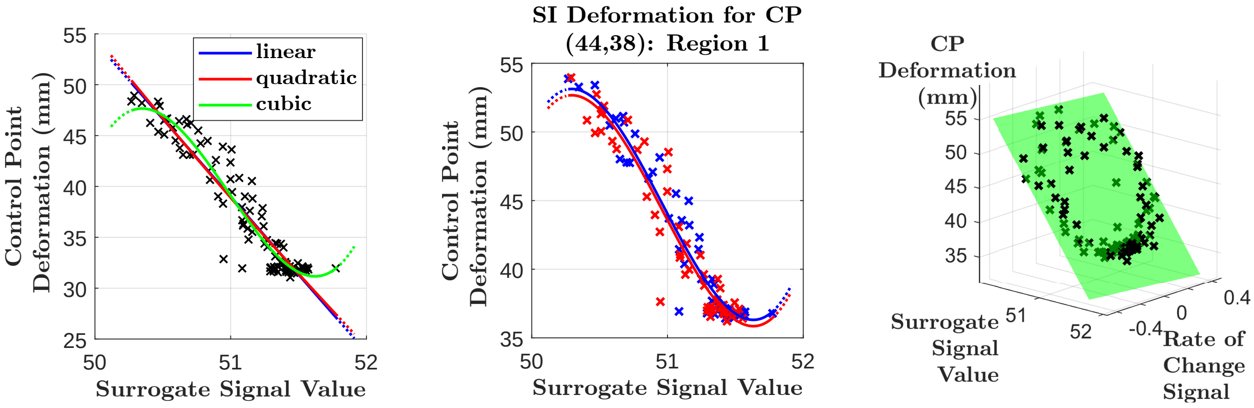
The performance of transformations from B-spline image registration [4] were first considered as they were used as a target for comparison of correspondence models. Visual inspection of an arbitrary target image, $\mathbf{I}_0$, and the corresponding deformed source image using the B-spline transformations, $\mathbf{I}_{\mathrm{BSp}}$, in Figure 1 indicates that image registration has performed well on the whole. The map of absolute intensity differences between the two in Figure 1c reveals slight differences at edges (e.g. between lungs and diaphragm, region 1 and region 2). Residual motion is to be expected due to quantisations involved with finite resolutions of the MR voxel size, slice thickness and control point spacing. Residual motion in the lateral-medial (LM) direction will also not be captured by the transformations – blood vessels and other small structures in the lung may therefore drift along this direction into and out of the 10 mm thick slice over the course of capture.

2.2 Correspondence Models

Table 1: Summary of correspondence models explored

Model type	Order (p)	# Surrogates (k)	Identifier	Details	Surrogates Explored
Baseline	$\{1,2,3\}$	1	$\mathbf{B}\langle p\rangle$ $\mathbf{C}\langle p\rangle$ $\mathbf{D}\langle p\rangle$	$\phi(s) = \sum_{i=0}^{p} c_i s^i$	Baseline (abdomen) Chest Diaphragm
Multiple Surrogates	1	$\{2, 5, 10,$ $15, 20, 25,$ $30, 50\}$	$\mathbf{M}\langle k\rangle$	$\phi(\mathbf{s}) = \sum_{i=1}^{k} c_i s_i + c_0$	Multiple (skin)
Split-cycle	$\{1,2,3\}$	1	$\mathbf{S}\langle p\rangle$	$\phi(s) = \sum_{i=0}^{p} c_i s^i$	Baseline (abdomen)

Surrogate + derivative	$\{1,2,3\}$	2	**2D-$\langle p \rangle$**	$\phi(s,s') = \sum_{i=0}^{p} \sum_{j=0}^{p-i} c_{i,j} s^i s'^j$	Baseline (abdomen)
Surrogate + derivative + 2nd derivative	1	3	**3D**	$\phi(s,s',s'') = c_3 s + c_2 s' + c_1 s'' + c_0$	Baseline (abdomen)



(a) (b) (c)

Figure 2: Example plot of model parameters against an arbitrary control point deformation (SI-direction) showing the results from the individual registrations (crosses) and the model fits: (a) Baseline 1D models, (b) split-cycle model with $p = 3$ (blue = inhale model, red = exhale model) and (c) 2D model. Dotted lines visualise extrapolated values.

The purpose of correspondence models is to learn a mapping ϕ from the surrogate signals $\mathbf{s}$ to a motion parameter $x = \phi(\mathbf{s})$. Motion parameters in this context are the deformations as described by each control point. A range of functions are considered and detailed in Table 1. At a high level, correspondence models can be split into:

- **Baseline**: 1D models consisting of a single surrogate with polynomial order 1 (linear), 2 (quadratic) or 3 (cubic). Such models constrain the motion parameters to follow a fixed trajectory regardless of breathing cycle, meaning it is not possible to model the hysteresis between inhalation and exhalation. They may be able to account for a small amount of inter-cycle variation e.g. different amplitudes of signal related to different depths of breathing (Figure 2a).
- **Multiple skin surface surrogates**: these are multivariate linear models where k surrogates are treated as independent variables. This allows increased flexibility to capture variations in breathing, but can easily lead to over-fitting [6] (see §3.3).
- **Split-cycle**: Same as the baseline 1D models, the crucial difference being that two separate models are fit to surrogate data depending on whether it is during inhalation or exhalation. This enables modelling of the intra-cycle variation as well as limited inter-cycle variation similar to the baseline (Figure 2b).
- **Signal + derivative**: 2D models consisting of a single surrogate and its time derivative. In the case of order $p = 1$, the motion parameters are constrained to lie on a 2D plane; having the time derivative of the surrogate as a second independent variable therefore allows intra-cycle variation to be modelled without requiring inhalation/exhalation split. There may also more capacity for modelling inter-cycle variation (Figure 2c).
- **Signal + derivative + second derivative**: A 3D linear model (only $p = 1$ considered) consisting of a single surrogate, its time derivative and its second time derivative. The second time derivative can be thought of as the 'acceleration' of the surrogate signal and therefore offers an extra degree of freedom to capture intra-cycle and inter-cycle variability.

2.3 Surrogate Signals

Two variants of image-derived scalar surrogate signals are considered in Figure 3a: skin surface displacement in the AP-direction and diaphragm displacement in the SI-direction, with respective datums of the left ($x = 0$) and top ($y = 0$) of the image frame. A threshold intensity of 20 is used for detecting the skin surface, and 10 for detecting the diaphragm. Models are primarily tested using the Baseline surrogate signal.

Where multiple skin surface displacements are required, they are spaced equally in the y- direction between chest ($y = 50$) and abdomen ($y = 150$) (Figure 3b). This range is purposely chosen as both chest and

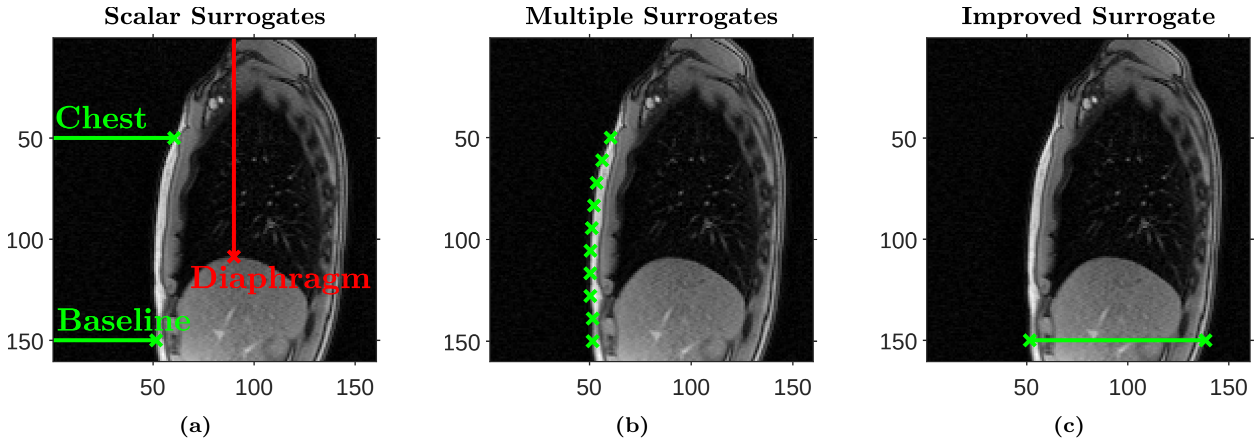


Figure 3: Image derived surrogate signals: (a) single scalar surrogates for skin surface and diaphragm surface displacements, (b) multiple surrogates on skin surface, (c) Proposal to improve the baseline surrogate.

abdomen signals could be complementary, helping to incorporate variation between thoracic (i.e. shallow) and abdominal breathing [8].

Deriving surrogate signals from internal imaging data may be susceptible to some issues. The first is that of signal accuracy; interpolation is required for the signal to be usable due to the relatively coarse in-slice resolution of 2 × 2mm^2, resulting in interpolation error when estimating the signal. Global translation of the subject over time may also cause signal drift if translation is parallel to the direction of measurement, or may result in the signal being generated from multiple skin/ diaphragm surface points if translation is perpendicular to the direction of measurement. A proposed solution to mitigate the effect of parallel translation could be to track the difference in skin surface position at the front and back of the subject, as shown in Figure 3c. In [10] the use of global surrogates, derived using PCA on image intensities, appears a promising alternative.

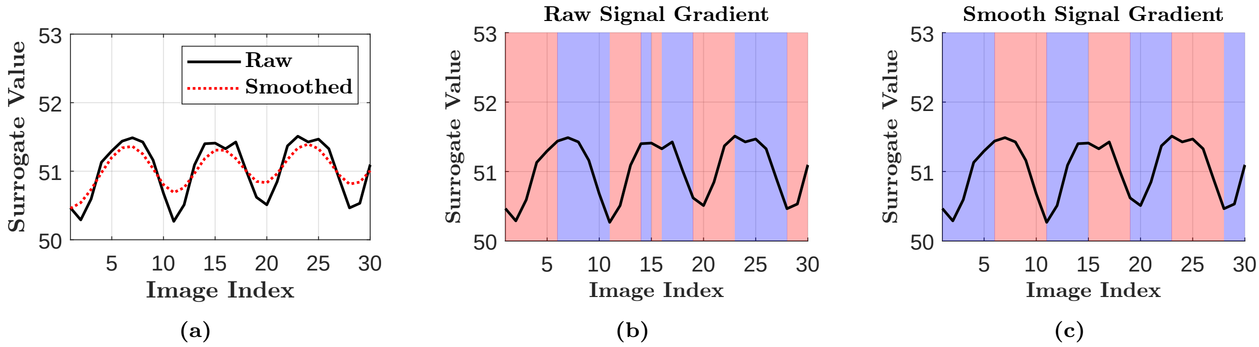


Figure 4: Determining inhalation (blue) and exhalation (red) phases of the breathing cycle: (a) Illustration of pseudo-Gaussian filtering on baseline surrogate, (b) phasing using gradient derived from raw surrogate, (c) phasing using gradient derived from surrogate smoothed with a pseudo-Gaussian filter.

Where time derivatives of surrogates are required either as features or to indicate inhalation or exhalation, a pseudo-Gaussian filter [7] is applied before gradient estimation. This reduces false detection of sign changes in the signal gradient due to higher frequency noise, improving the partitioning as seen in Figure 4.

2.4 Fitting Procedure

Model coefficients, $\mathbf{c}$ are determined via ordinary least squares. For T time points (images), M motion parameters, and n model coefficients, this can be written in matrix notation as $\mathbf{X} = \mathbf{SC}$, where $\mathbf{X} \in \mathbb{R}^{T\times M}$ is a matrix of motion parameters, $\mathbf{S} \in \mathbb{R}^{T\times n}$ a matrix of surrogate values, and $\mathbf{C} \in \mathbb{R}^{n\times M}$ a matrix of model coefficients. The solution to the least squares objective, $\operatorname{argmin}_{\mathbf{C}} (\mathbf{X} - \mathbf{SC})^2$, can be found via the Moore-Penrose matrix inverse:

$$\mathbf{C} = \left(\mathbf{S}^{\mathrm{H}}\mathbf{S}\right)^{-1} \mathbf{S}^{\mathrm{H}}\mathbf{X}.$$

The first 100 training images are used to fit the models with the remaining 1400 for testing. This is representative of situations where the correspondence model is used for prediction; the data is treated as a time-series and splitting in this manner ensures that look-ahead bias is avoided. The 100 images represent roughly 33 s of data capture and 12 cycles of breathing so intra-cycle variation will be present in the training data. Inter-cycle variation may be limited and increasing the size of the training set should capture more of the variation and lead to improved generalisation performance.

Cross-validation techniques could be considered an improvement over the single train/test split method as it uses the data more completely to correct for any selection bias that stems from the choice of split. They also provide multiple (e.g. k in k-fold cross-validation) independent measures of error that can be averaged to provide a more robust estimate [3]. Modification of these techniques are required as independence of the observations does not hold for time series data [2] and often the assumption of stationarity also does not. An overview of such methods is detailed in [3].

When using a large number of skin surface surrogates (e.g. 50), Principal Components Regression (PCR) [6] is applied to determine whether it helps prevent over-fitting. Here, Principal Components Analysis (PCA) is applied to the surrogate data and the data transformed to a lower-dimensional representation that maximises explained variance. This lower-dimensional representation is then used as the inputs to the model.

2.5 Model Assessment

2.5.1 Evaluation of Fitting Procedure

The following metrics are based on the training data alone. In the formulae below, x is a single motion parameter, $\hat{x}$ is the estimate of the motion parameter using the correspondence model, T is the number of training samples and N is the number of model parameters to be estimated (including implicit estimate of variance). The assumption of Gaussian noise allows use of a simplified term for the likelihood for AIC and BIC [1]:

1. **Mean Absolute Error, MAE** $= \frac{1}{T}\sum_{i=1}^{T}|x_i - \hat{x}_i|$

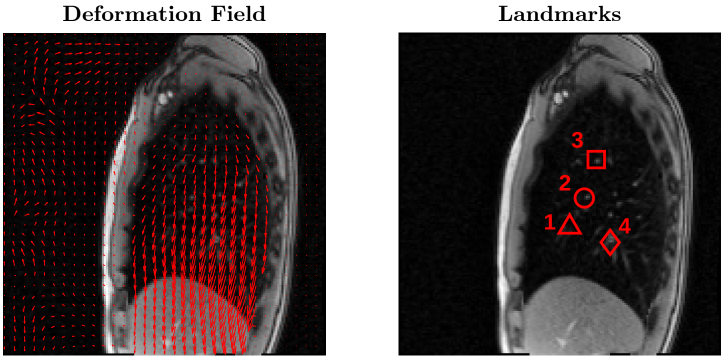
2. **Root Mean Square Error, RMSE** $= \sqrt{\frac{1}{T}\sum_{i=1}^{T}(x_i - \hat{x}_i)^2}$

3. $\mathbf{AIC_C} = 2N + T\log\left(\frac{1}{T}\sum_{i=1}^{T}(x_i - \hat{x}_i)^2\right) + \frac{2N(N+1)}{T-N-1}$

4. $\mathbf{BIC} = N\log T + T\log\left(\frac{1}{T}\sum_{i=1}^{T}(x_i - \hat{x}_i)^2\right)$

These statistics are evaluated for every motion parameter and summary statistics (mean, standard deviation etc.) calculated.

2.5.2 Evaluation Metrics on Test Set


Deformation Field

Landmarks

(a)

(b)

Figure 5: Example images at end-inhalation showing: (a) deformation field from B-spline registration, (b) landmark positions.

The following metrics are calculated using the test set not involved in fitting the models. These give a better indication of model performance when prediction is the focus.

1. **Visual Assessment:** This follows the procedure outlined in §2.1.1, where maps of the absolute intensity difference between images deformed by model-estimated transformations and images deformed by B-spline registrations were produced, $\mathbf{\Delta I} = |\mathbf{I}_{\text{est}} - \mathbf{I}_{\text{BSp}}|$. Due to the subjective nature of assessing these maps, accompanying summary statistics – including the correlation coefficient between the images – were produced to quantify what was seen visually.

2. **Deformation Field Error (DFE):** The deformation field describes how every voxel in the source ('static') image is mapped to a corresponding voxel in the target ('moving') image as a relative displacement [9]. The nature of this displacement is best visualised with an end-inhalation image as in Figure 5a. The quality of the model-estimated deformation field was evaluated by calculating the Euclidean (L2-norm) difference to that from the B-spline registration at every voxel *within the body*. The deformation field error gives good indication of how the model performs across the entire region.

A drawback of reporting only the L2-norm error is not understanding which component – AP or SI – has more error.

3. **Landmark Error (LME):** Four blood vessels marked in Figure 5b were tracked in all target images to provide 'ground truth' co-ordinates. The model-estimated deformation fields are then applied to this landmark co-ordinates to transform them into the space of the source image, where the AP-, SI- and L2-norm error in position can then be calculated relative to the ground truth position in the source image. This is useful when specific regions and/or features of interest are the focus. However, the labelling of ground truth positions itself is done automatically and may have its own associated error to the true position of the landmarks.

2.5.3 Parameter Uncertainty

Uncertainty in baseline model parameters for the SI-component of CP(44,38) were estimated by constructing the posterior distribution of the parameters. Due to the dependence of the data (time series), the residual bootstrap was selected to preserve this structure. $10,000$ bootstrapped iterations were fit and the resulting posterior distributions can be found in Figure 6.

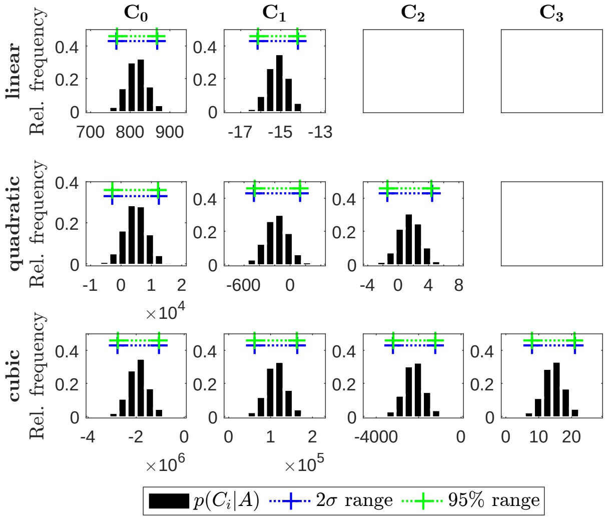


Figure 6: Posterior distributions of parameters for baseline models estimated using residual bootstrap.

The 2σ and 95% ranges on all parameters are roughly equal, indicating the posterior distributions are approximately Gaussian. Where the models differ is that the uncertainty range in common parameters (C_0, C_1) are orders of magnitude larger as model order increases.

3 Results & Discussion

Table 2: Model Fitting Evaluation - Training Set Only. For split models, reported AIC_C and BIC are average of the inhale and exhale models.

	MAE				RMSE					
Model ID	Mean	σ	95th percentile	Max	Mean	σ	95th percentile	Max	AIC_C $(\times 10^2)$	BIC $(\times 10^2)$
B1	0.41	0.48	1.62	2.39	0.55	0.64	2.17	3.12	-1.875	-1.800
B2	0.40	0.47	1.57	2.33	0.54	0.62	2.12	3.04	-1.884	-1.784
B3	0.38	0.42	1.43	2.03	0.52	0.58	2.00	2.82	-1.893	-1.769
C1	0.75	1.17	3.72	5.98	0.95	1.43	4.58	7.22	-1.586	-1.510
C2	0.69	1.05	3.34	5.32	0.89	1.32	4.22	6.67	-1.623	-1.523
C3	0.69	1.04	3.33	5.30	0.89	1.32	4.22	6.67	-1.636	-1.512
D1	0.31	0.29	0.87	2.01	0.42	0.41	1.19	2.83	-1.989	-1.914
D2	0.29	0.27	0.83	2.03	0.41	0.39	1.16	2.82	-2.010	-1.910
D3	0.29	0.26	0.81	1.83	0.40	0.38	1.15	2.74	-2.002	-1.878
M2	0.41	0.48	1.58	2.37	0.55	0.63	2.14	3.08	-1.907	-1.807
M5	0.39	0.47	1.55	2.35	0.53	0.62	2.11	3.03	-1.929	-1.758
M10	0.35	0.40	1.36	2.02	0.47	0.52	1.81	2.64	-1.963	-1.687
M50	0.25	0.26	0.91	1.51	0.32	0.34	1.17	1.97	-0.749	-0.568
M5(PCR)	0.40	0.48	1.59	2.41	0.53	0.63	2.14	3.09	-1.925	-1.755
S1	0.40	0.47	1.59	2.32	0.53	0.62	2.12	3.05	-0.946	-0.894
S2	0.38	0.45	1.54	2.24	0.51	0.60	2.06	2.98	-0.953	-0.886
S3	0.35	0.39	1.36	1.94	0.49	0.55	1.91	2.71	-0.954	-0.872
2D-1	0.39	0.47	1.59	2.31	0.53	0.61	2.12	3.00	-1.915	-1.815
2D-2	0.37	0.43	1.50	2.09	0.51	0.58	2.02	2.87	-1.929	-1.758
2D-3	0.34	0.38	1.33	2.00	0.47	0.53	1.86	2.77	-1.907	-1.650

3D	0.37	0.42	1.42	2.24	0.49	0.55	1.82	2.85	-1.957	-1.834

Table 3: Summary statistics of visual assessment based on absolute intensity difference across all voxels and all test images.

Model ID	Mean	σ	95th percentile	Corr
B1	1.23	2.35	4.07	0.981
B2	1.25	2.37	4.13	0.981
B3	1.15	2.15	3.75	0.984
C1	1.23	2.24	4.10	0.982
C2	1.34	2.62	4.50	0.977
C3	1.41	2.83	4.84	0.974
D1	0.93	1.37	3.06	0.993
D2	0.92	1.38	3.02	0.993
D3	0.90	1.33	2.95	0.993
M2	1.22	2.34	4.02	0.982
M5	1.14	2.22	3.68	0.983
M10	1.19	2.26	3.87	0.983
M50	1.26	2.27	4.04	0.982
M5(PCR)	1.11	2.07	3.56	0.985
S1	1.23	2.34	4.05	0.981
S2	1.25	2.37	4.11	0.981
S3	1.22	2.29	4.00	0.982
2D-1	1.23	2.34	4.05	0.982
2D-2	1.23	2.34	4.08	0.982
2D-3	1.17	2.24	3.86	0.983
3D	1.02	1.66	3.36	0.990

Table 4: Summary statistics of L2-norm Deformation Field Error (mm) across all voxels within the body and all test images.

Model ID	Mean	σ	95th percentile
B1	2.33	2.69	8.16
B2	2.14	2.40	7.29
B3	1.84	2.35	7.02
C1	1.92	2.18	6.52
C2	2.65	4.33	9.96
C3	2.87	4.91	10.72
D1	0.87	0.69	2.05
D2	0.85	0.68	2.03
D3	0.82	0.67	1.97
M2	2.32	2.69	8.13
M5	2.10	2.51	7.54
M10	2.37	2.89	8.52
M50	2.48	2.94	8.76
M5(PCR)	1.96	2.29	6.90
S1	2.33	2.70	8.18
S2	2.14	2.42	7.37
S3	1.98	2.48	7.34
2D-1	2.32	2.68	8.15
2D-2	2.08	2.36	7.20
2D-3	2.00	2.63	7.65
3D	1.52	1.67	5.03

Table 5: Summary statistics of landmark errors (mm) - reported as AP and SI components as well as the L2-norm - across all test images. Landmark errors for B-spline transformations also reported as the 'gold' standard.

	AP				SI				L2-norm			
Model ID	Mean	σ	95th percentile	Max	Mean	σ	95th percentile	Max	Mean	σ	95th percentile	Max
BSpline	0.63	0.50	1.52	3.53	0.82	0.68	2.20	4.66	1.12	0.74	2.58	5.72
B1	1.01	0.81	2.55	4.00	3.11	2.37	8.17	13.81	3.46	2.23	8.22	13.86
B2	1.04	0.82	2.59	3.98	3.01	2.33	7.97	13.84	3.39	2.19	8.01	13.89
B3	0.98	0.79	2.54	4.02	2.25	2.26	7.08	14.92	2.68	2.13	7.15	15.01
C1	1.21	0.85	2.80	4.15	2.33	1.95	6.08	13.84	2.86	1.79	6.27	14.39
C2	1.25	1.01	3.06	7.53	3.50	4.19	11.68	31.46	3.97	4.07	11.76	31.88
C3	1.27	0.98	3.02	6.83	3.83	5.07	13.55	39.56	4.30	4.94	13.67	39.98
D1	1.26	0.90	2.87	4.48	1.00	0.68	2.28	5.98	1.72	0.94	3.38	6.73
D2	1.26	0.89	2.90	4.54	1.01	0.71	2.39	6.22	1.73	0.97	3.46	6.99
D3	1.21	0.89	2.95	4.30	0.99	0.71	2.37	5.79	1.67	0.96	3.53	6.26
M2	1.02	0.80	2.61	4.09	3.11	2.37	8.17	13.82	3.47	2.22	8.22	13.85
M5	0.83	0.57	1.90	3.82	2.70	2.19	7.33	13.15	2.95	2.10	7.42	13.30
M10	0.91	0.69	2.27	3.94	2.95	2.32	7.72	15.17	3.24	2.21	7.80	15.22
M50	1.23	0.92	2.94	6.05	2.88	2.23	7.67	14.30	3.35	2.10	7.76	14.30
M5(PCR)	0.95	0.71	2.22	3.87	2.53	2.03	6.77	12.69	2.87	1.92	6.82	12.73
S1	1.00	0.75	2.40	4.01	3.12	2.36	8.14	13.82	3.46	2.21	8.21	13.88
S2	1.12	0.84	2.69	5.36	3.04	2.35	8.03	13.86	3.43	2.22	8.14	13.90
S3	1.41	1.19	3.59	11.44	2.30	2.28	7.15	14.45	3.01	2.20	7.55	17.83
2D-1	0.96	0.74	2.38	3.88	3.10	2.35	8.09	13.63	3.43	2.20	8.12	13.74
2D-2	1.00	0.77	2.43	3.63	2.96	2.32	7.90	13.40	3.31	2.18	7.93	13.50
2D-3	1.01	0.77	2.44	6.99	2.42	2.38	7.55	20.58	2.85	2.23	7.63	21.73
3D	1.07	0.75	2.48	3.92	1.70	1.37	4.45	9.49	2.21	1.25	4.56	9.49

3.1 Baseline Models

Baseline 1D models and surrogate (B1-3) were fit according to §2.4. As expected, the residual fitting error reported in terms of MAE and RMSE in Table 2, decreases with increasing polynomial order. From a model selection perspective, there is disagreement between the two information criterions; the cubic model has the lowest AIC_C whereas the linear model has the lowest BIC.

Absolute intensity differences between model-deformed images and registration-deformed images were assessed visually in Figure 7. Performance is best for early test images, but degrades for later test images (worst around the middle of test images); the reasoning is discussed further in §3.6. Across the three windows of test data, the cubic model depicts lower global intensity difference compared to the linear and

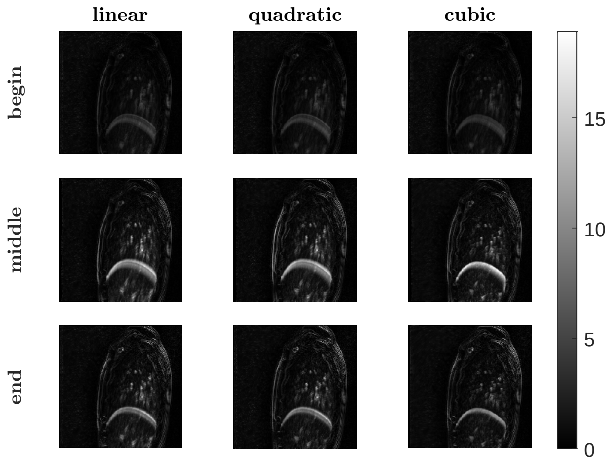


Figure 7: Absolute intensity difference maps between source images deformed by B-spline registration and (baseline) model-estimated transformations. Results averaged over 200 images: begin = 1-200, middle = 801-1000, end = 1201-1400.

quadratic models. However, locally the cubic model does show higher intensity differences around the lung-diaphragm boundary for the middle test images. This is supported from the summary statistics in Table 3 with the cubic model having lower mean absolute intensity difference and higher correlation, but higher maximum absolute intensity difference.

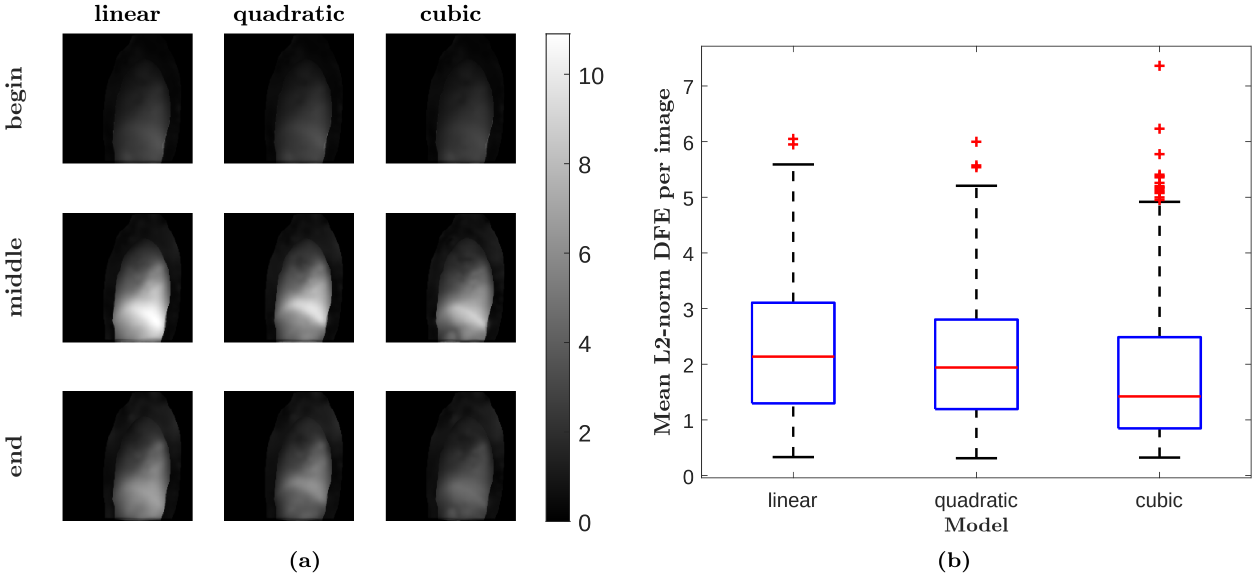


Figure 8: L2-norm Deformation Field Error (mm) for baseline models: (a) maps across different windows of test images, (b) boxplot showing distribution of mean error over all test images.

Analysis of the DFE (Table 4) tells a similar story. DFE maps in Figure 8a show lower global error for the cubic model across all time windows while the boxplots in Figure 8b highlights that, although the cubic model has lower overall error, it can occasionally produce a higher error. The DFE maps show a similar pattern of performance degradation over time.

Landmark error statistics in Table 5 provide further evidence that the cubic model is the best baseline with both AP and SI components showing the lowest mean error. There is also less spread in the error as measured by the standard deviation and a lower 95th percentile error. Figure 9 breaks down the L2-norm error per landmark; this is important as landmark 4 exhibits significantly higher error, therefore summary statistics may be heavily biased by one landmark. The cubic model has lower error across all four landmarks but again, can produce the highest landmark errors on occasion.

Polynomial models of the surrogate signal appear to capture more of the internal motion, but over-fitting creates a higher risk of extrapolation errors. Solutions to this are reviewed in [6] e.g. by reverting back to a linear model when values of surrogate signal not seen during training are encountered.

3.2 Alternative Surrogates

1D models for skin surface displacement at the chest (C1-3) and diaphragm (D1-3) were compared to the baseline abdomen surrogate. The linear model with a chest surrogate outperforms the linear and quadratic baseline models across all evaluation test metrics, despite having poorer statistics on the training data.

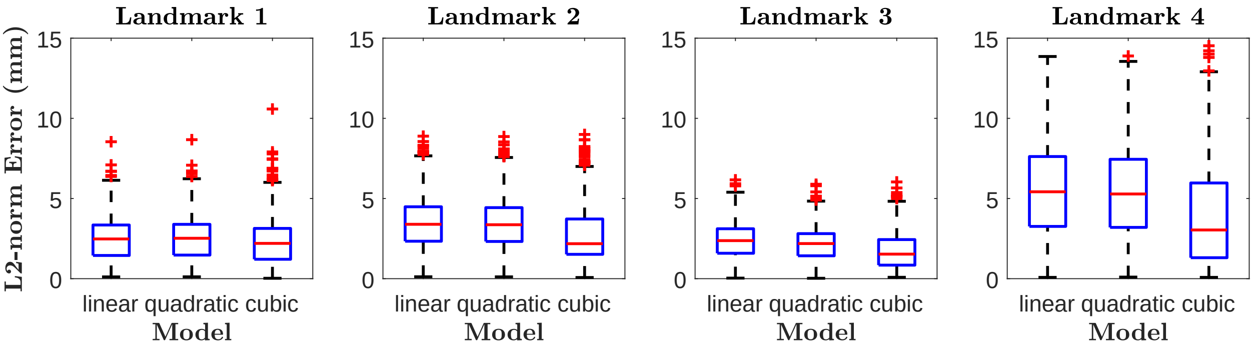
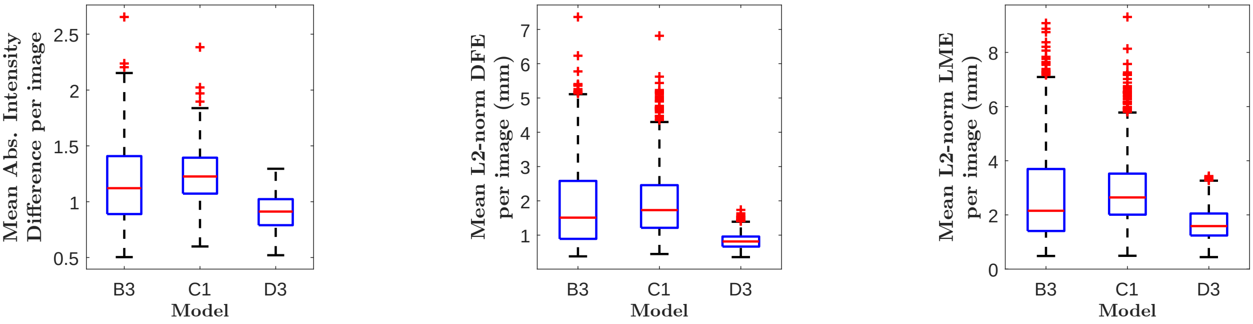


Figure 9: Boxplots showing the distribution of landmark errors (L2-norm) for each of the four landmarks over all test images.

The diaphragm surrogate shows vast improvement in all areas of evaluation, with the linear model outperforming all models of skin surface surrogates investigated. This is likely because diaphragm movement has a stronger relationship to lung volume compared to skin movement. There may also be less drift of the subject in the SI direction over the data set compared to AP direction, making it easier to predict. The best 1D models using each of the surrogates are compared in Figure 10.



(a) **(b)** **(c)**

Figure 10: Comparison of evaluation metrics on best 1D models for different surrogates: (a) Absolute Intensity Error, (b) Deformation Field Error, (c) Landmark Error.

3.3 Multiple Surrogates

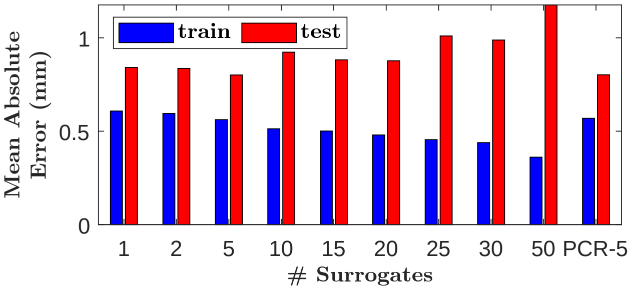


Figure 11: Evolution of mean absolute error for train and test set with different numbers of skin surface surrogates.

A combination of up to 50 skin surface surrogates were used and the (mean) MAE calculated from both the training and test set are plotted in Figure 11. Tables 3 to 5 indicate slight improvement in model performance up to 5 surrogates. Increasing the number of surrogates further results in a divergence in the residual fitting error between training and test set – an indication that over-fitting to the training data is occurring. To resolve the issue of over-fitting, PCA is applied to reduce the dimensionality of the surrogate data from 50 signals to just 5 before fitting the model M5(PCR). The gap in MAE between training and test set for this model is reduced by doing so, and slight improvements to evaluation metrics can be seen over the naive choice of 5 surrogate signals. However, it is not clear whether this is by chance or if the application of PCA has really found more meaningful surrogate signals.

Success of PCR on such data might be limited as it does not take the structure of the internal motion data into account [11]. Partial Least Squares Regression is a supervised technique of dimensionality reduction that seeks to resolve this issue and may be an avenue to explore in the future.

3.4 Modelling Inhalation and Exhalation Separately

Fewer training images are available when building separate inhalation and exhalation models as they must be split according to the cycle. There is an expected improvement in residual fitting error on the training set but not in performance on the test set – for the quadratic and cubic models (S2, S3), performance degrades relative to the baseline models across all three assessments. An increased likelihood of over-fitting due to fewer training images (per model) may be partially responsible for this. Discontinuities when switching between the models may also result in larger errors.

Improvements are suggested in [5], where end-cycle data is used in *both* inhalation and exhalation models and equality constraints imposed so that inhalation and exhalation curves meet at end-cycle positions.

3.5 Surrogates + time derivatives

Utilising the 1st time derivative shows minor improvement in evaluation metrics (Tables 3 to 5) versus surrogate-only baseline counterparts for 1st- and 2nd-order models. The 3rd-order model (2D-3), while having better performance compared to 2D-1 and 2D-2, does not outperform its counterpart 1D baseline (B3). It is likely that 2D-3 has over-fit the training data due to having 11 model parameters. Interestingly, increasing polynomial order lowers the SI component of landmark error but increases the AP component for the 2D models.

Including the second time derivative of the baseline surrogate sees a substantial improvement across all evaluation metrics; this is the best performing model found for skin surface surrogates and it is assumed important information is encoded in the surrogate acceleration. Performance relative to the best baseline model (B3) and best overall model (D3) is summarised in Figure 12. The diaphragm surface surrogate still offers the best results.

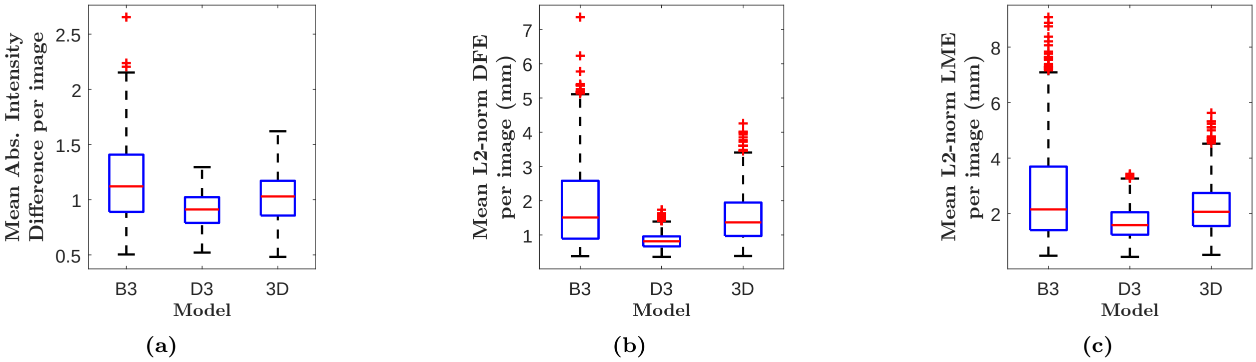


Figure 12: Comparison of 3D model (3D) against best baseline model (B3) and best overall model (D3): (a) Absolute Intensity Error, (b) Deformation Field Error, (c) Landmark Error.

3.6 Surrogate Signal Drift

Stability in the surrogate signal(s) over time is central to maintaining good model performance. For this data set, drift in the baseline surrogate signal level and amplitude (Figures 13a and 14a) causes degradation in the majority of models, particularly those that rely on the surrogate value alone. The 3D model fares better against surrogate drift as the derived rate of change and acceleration signals seen during training (Figures 13b and 13c) are representative of that seen in the test set.

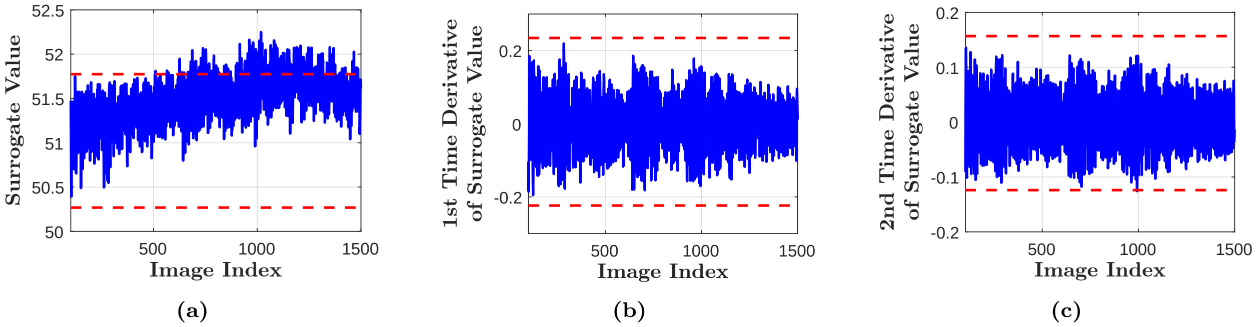
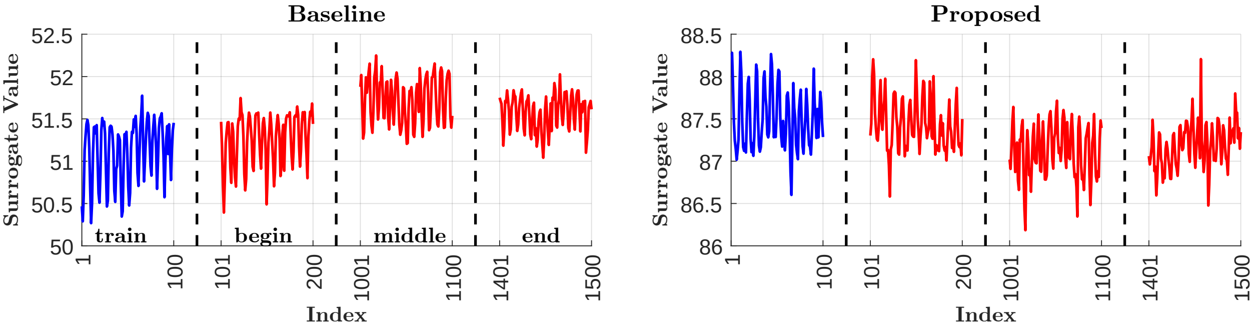


Figure 13: Visualisation of surrogate signals during test (blue) relative to limits seen during training (red dotted lines): (a) surrogate value, (b) its 1st time derivative, (c) its 2nd time derivative.

It is calculated that roughly 20% of the surrogate values encountered during testing are outside of the range seen during training. Moving to a proposed surrogate as described in Figure 3c shows a reduced (but not eliminated) drift over the test set in Figure 14b. Only $1 - 2\%$ of values encountered during testing lie outside of the training range for this surrogate.



(a) (b)

Figure 14: Evolution of surrogate signals (blue = train, red = test): (a) baseline surrogate, (b) proposed surrogate. The baseline signal shows both drift in level and reduced amplitude at the end of test.

4 Conclusion

Numerous correspondence models have been presented for estimating the internal motion using a variety of image-derived surrogate signals. Higher order models appear to better estimate the internal motion but may risk over-fitting with occasional larger extrapolation errors. Incorporating 1st and 2nd time-derivative data shows the most promise for skin surface surrogates as it enables flexibility to model both intra-cycle and inter-cycle variation, whilst capturing important information that better generalises to unseen data.

Models that incorporate many skin surface surrogates can theoretically capture more variation in the internal motion, but require dimensionality reduction or regularisation techniques to prevent over-fitting. Application of PCR still does not prove as successful as using a diaphragm surrogate. This presents an ongoing challenge as skin surrogates are much more easily and cheaply obtained.

It is important to consider how surrogate generation can be made more robust to reduce the amount of temporal change to the signal where feasible. Use of global surrogates instead of local surrogates as mentioned in [10] may provide one such way to do this.

References

[1] Daniel Alexander. *Computational Modeling for Biomedical Imaging: COMP0118 Notes.* 2020.

[2] S. Arlot and Alain Celisse. “A survey of cross-validation procedures for model selection”. In: *Statistics Surveys* 4 (2010), pp. 40–79.

[3] Christoph Bergmeir and José M. Benítez. “On the use of cross-validation for time series predictor evaluation”. In: *Information Sciences* 191 (2012). Data Mining for Software Trustworthiness, pp. 192–213. ISSN: 0020-0255. DOI: https://doi.org/10.1016/j.ins.2011.12.028. URL: https://www.sciencedirect.com/science/article/pii/S0020025511006773.

[4] B. Eiben et al. “Statistical Motion Mask and Sliding Registration”. In: *WBIR.* 2018.

[5] A.P. King et al. “A subject-specific technique for respiratory motion correction in image-guided cardiac catheterisation procedures”. In: *Medical Image Analysis* 13.3 (2009), pp. 419–431. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2009.01.003. URL: https://www.sciencedirect.com/science/article/pii/S1361841509000048.

[6] J.R. McClelland et al. “Respiratory motion models: A review”. In: *Medical Image Analysis* 17.1 (2013), pp. 19–42. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2012.09.005. URL: https://www.sciencedirect.com/science/article/pii/S136184151200134X.

[7] Tom O’Haver. *Fast smoothing function.* 2021. URL: https://www.mathworks.com/matlabcentral/fileexchange/19998-fast-smoothing-function.

[8] Freddy Odille et al. “Generalized MRI reconstruction including elastic physiological motion and coil sensitivity encoding”. In: *Magnetic Resonance in Medicine* 59.6 (2008), pp. 1401–1411. DOI: https://doi.org/10.1002/mrm.21520. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mrm.21520. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.21520.

[9] James Shackleford, Nagarajan Kandasamy, and Gregory Sharp. *High Performance Deformable Image Registration Algorithms for Manycore Processors.* 1st. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2013. ISBN: 0124077412.

[10] Elena H Tran et al. “Evaluation of MRI-derived surrogate signals to model respiratory motion”. In: *Biomed Phys Eng Express* 6 (2020). DOI: https://doi.org/10.1088/2057-1976/ab944c.

[11] M Wilms et al. “Multivariate regression approaches for surrogate-based diffeomorphic estimation of respiratory motion in radiation therapy”. In: *Physics in Medicine and Biology* 59 (2014), p. 1147.