# Surrogate-Driven Estimation of Respiratory Motion and Layers in X-Ray Fluoroscopy

Peter Fischer[1], Thomas Pohl[2], Andreas Maier[1], and Joachim Hornegger[1]

[1] Pattern Recognition Lab and Erlangen Graduate School in Advanced Optical Technologies (SAOT), FAU Erlangen-Nürnberg, Erlangen, Germany,
peter.fischer@fau.de
[2] Siemens Healthcare, Forchheim, Germany

**Abstract.** Dense motion estimation in X-ray fluoroscopy is challenging due to low soft-tissue contrast and the transparent projection of 3-D information to 2-D. Motion layers have been introduced as an intermediate representation, but so far failed to generate plausible motions because their estimation is ill-posed. To attain plausible motions, we include prior information for each motion layer in the form of a surrogate signal. In particular, we extract a respiratory signal from the images using manifold learning and use it to define a surrogate-driven motion model. The model is incorporated into an energy minimization framework with smoothness priors to enable motion estimation.
Experimentally, our method estimates 48% of the 2-D motion field on XCAT phantom data. On real X-ray sequences, the target registration error of manually annotated landmarks is reduced by 52%. In addition, we qualitatively show that a meaningful separation into motion layers is achieved.

## 1 Introduction

X-ray fluoroscopy is an important modality for guidance of minimally-invasive interventions. It has good spatial and temporal resolution and clearly visualizes interventional devices and bones. However, the contrast of soft tissue is low and 3-D information is lost due to the transparent projection to 2-D. In this paper, we deal with dense motion estimation in X-ray images. There are many clinical applications of fluoroscopy for which this is beneficial. Temporal denoising algorithms depend on accurate motion estimates [1]. Coronary DSA requires the compensation of cardiac and respiratory motion occurring between the mask and the contrasted image [14]. In thoracic and abdominal interventions, fusion of X-ray images with previously acquired roadmap overlays, created from contrasted images, CT, MR, or C-arm CT, requires motion compensation to correctly display the overlays on the live fluoroscopic images [4,15].

The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-319-24553-9_35

There are two major challenges for motion estimation in fluoroscopy. First, the low soft-tissue contrast complicates intensity-based image registration, because common similarity measures are dominated by high-contrast structures. Second, the estimation of 2-D motion in X-ray fluoroscopy is ill-posed due to the transparent projection of differently moving 3-D structures to 2-D. To alleviate the transparency problem, motion layers have been introduced [11]. The goal of motion estimation is then to compute a separate 2-D motion field for each layer.

Two approaches have been proposed for this problem. The first approach avoids to compute motion layers and directly estimates multiple 2-D motions for each pixel. Assuming locally constant motion in space and time in each layer, the common brightness constancy assumption can be extended to the transparent setting [1,11]. Alternatively, certain special motion types can be estimated in transparency, e.g., parametric motions [3] or a single non-static layer [15]. However, these assumptions are restrictive and cumbersome for more than two layers. The second approach is estimation of layers and motions. This leads to a chicken-and-egg problem, i.e., it is easy to compute the layers when the motion is known and vice versa. Szeliski et al. assume parametric motion to simplify the problem [12]. Preston et al. introduce a layer gradient penalty for the layers and a smoothness prior for the motions [9]. However, the motions and layers are not physiologically meaningful, restricting their usefulness to applications where the layers and motions are recombined, e.g., frame interpolation or denoising. Surrogate signals are commonly used in respiratory motion models [8]. For example, Martin et al. estimate a surrogate-driven 3-D motion field in motion-compensated C-arm CT reconstruction [7] .

In this work, we propose to enhance layered motion estimation using a separate surrogate signal for each layer. In particular, we use a static layer without motion and a respiratory layer with motion proportional to a respiratory surrogate signal. The respiratory surrogate signal is extracted from the X-ray sequence using manifold learning. Together with smoothness priors for layers and motions, this enables us to retrospectively estimate physiologically plausible layers and motions in an energy formulation. The proposed method is especially suited for building 2-D respiratory motion models, as the dependence of the motion on the surrogate signals is required anyways. We quantify the estimation error of the proposed method on simulated X-ray images by comparing the estimated to the ground truth motion. On clinical sequences, we evaluate quantitatively using manual annotations and qualitatively show that the respiratory motion is accurately captured and that the motion layers are separated.

## 2 Methods

### 2.1 Image Formation Model

We are interested in separating X-ray images $I(\boldsymbol{x}, t), t \in \{1, \ldots, T\}$ into different motion layers $L_l(\boldsymbol{x})$, where each layer may undergo independent non-parametric 2-D motion $\boldsymbol{v}_l(\boldsymbol{x}, t) \in \mathbb{R}^2$. $\boldsymbol{x} \in \mathbb{R}^2$ is the image pixel position. A motion layer can roughly be assigned to each source of motion, e.g., breathing, heartbeat, and

background. The images are created additively from the transformed layers as

$$I(\boldsymbol{x}, t) = \sum_{l=1}^{N} L_l(\boldsymbol{x} - \boldsymbol{v}_l(\boldsymbol{x}, t)) + \eta \quad , \tag{1}$$

where $\eta \in \mathbb{R}$ is introduced to account for model errors and observation noise in the log-transformed X-ray model [9]. In this paper, we restrict the number of layers in the image sequence to $N = 2$, a static and a respiratory layer.

### 2.2 Surrogate-Driven Motion Model

The surrogate-driven model for layered motion is defined as

$$\boldsymbol{v}_l(\boldsymbol{x}, t) = s_l(t) \cdot \boldsymbol{\nu}_l(\boldsymbol{x}) \quad , \tag{2}$$

where $s_l(t) \in \mathbb{R}$ is a surrogate signal that is used to scale the base motion $\boldsymbol{\nu}_l(\boldsymbol{x}) \in \mathbb{R}^2$. The surrogate-driven motion model is crucial to achieve physiologically plausible motions. It reduces the number of motion parameters by a factor of $T-1$ compared to unconstrained motion fields, because $\boldsymbol{\nu}_l(\boldsymbol{x})$ is defined only for one point in time and extended to other times using Eq. (2), whereas unconstrained motion fields are defined for all points in time except $t = 0$. Thus, the parameter space is constrained to the subspace where the motion fields agree with the surrogate signals and thus with the underlying physiological processes.

In our application, the static layer with a constant surrogate signal $s_1(t) = 0$ is required to describe the static components of the X-ray sequence. The respiratory surrogate signal $s_2(t)$ can in principle be acquired by any means, e.g., spirometry or respiratory belt. In this work, we derive the respiratory signal directly from the intensities of the X-ray images using manifold learning. It has proven to be effective for X-ray fluoroscopy [5]. The advantage for our application is that the signal is based on the same images that are used for motion estimation, thus facilitating the proportionality assumption in Eq. (2).

### 2.3 Motion and Layer Estimation

To define a tractable optimization problem for joint motion and layer estimation, we include Eq. (1) and Eq. (2) into an energy formulation

$$E(\mathcal{L}, \mathcal{V}) = D(\mathcal{L}, \mathcal{V}) + \lambda_L R(\mathcal{L}) + \lambda_\nu R(\mathcal{V}) \quad , \tag{3}$$

where $D(\mathcal{L}, \mathcal{V})$ is the data term, $R(\mathcal{L})$ and $R(\mathcal{V})$ are regularization terms for the layers and motions and $\lambda_L, \lambda_\nu \in \mathbb{R}$ are their weights. $\mathcal{L}$ is the set of all layers $L_l$ and $\mathcal{V}$ is the set of all base motions $\boldsymbol{\nu}_l$.

The data term penalizes deviations from Eq. (1). Since the image formation model is only an inaccurate representation of the true X-ray generation process, robustness to outliers in the data term

$$D(\mathcal{L}, \mathcal{V}) = \sum_{t=1}^{T} \int_{\Omega} \psi\left( I(\boldsymbol{x}, t) - \sum_{l=1}^{N} L_l(\boldsymbol{x} - \boldsymbol{v}_l(\boldsymbol{x}, t)) \right) \mathrm{d}\boldsymbol{x} \tag{4}$$

is mandatory, where $\Omega$ is the image domain and $\psi$ is a robust penalty function. We use the Charbonnier penalty $\psi(z) = \sqrt{(\epsilon^2 + z^2)} - \epsilon$ with $\epsilon = 0.01$ as a differentiable approximation of the $L_1$-norm.

The regularization term for the layers is designed to favor spatially smooth layers, while still allowing for edges. Similar to denoising and reconstruction, we use a differentiable approximation of the isotropic total variation (TV) regularization

$$R(\mathcal{L}) = \sum_{l=1}^{N} \int_{\Omega} \psi\left(\|\nabla L_l(\boldsymbol{x})\|_2\right) \mathrm{d}\boldsymbol{x} \ , \tag{5}$$

where $\nabla = (\partial_x, \partial_y)^T$ is the spatial gradient. Preston et al. adapt the TV regularization to the image gradients [9]. In our experience, this does not improve the results, because Eq. (5) is already edge-preserving, and is much more expensive to compute, because all images must be warped to $t = 1$.

A similar spatial smoothness constraint is employed for the motions

$$R(\mathcal{V}) = \sum_{l=1}^{N} \int_{\Omega} \psi\left(\sqrt{\|\nabla \nu_l^x(\boldsymbol{x})\|_2^2 + \|\nabla \nu_l^y(\boldsymbol{x})\|_2^2}\right) \mathrm{d}\boldsymbol{x} \ , \tag{6}$$

where $\nu_l^x, \nu_l^y$ are the horizontal and vertical motions, respectively. Here, the robust penalty allows for motion boundaries. In addition, it is computationally less expensive than regularizing a full 2-D+t motion field for each layer, because it must be computed only for one point in time. In general, motion estimation benefits from regularization along the time-domain. However, this is already covered by the surrogate-driven motion model. In this sense, Eq. (2) is a strong regularization of the motion field along the surrogate signal.

### 2.4 Implementation

As the manifold learning method to extract the respiratory signal from the intensities of the entire X-ray image, we use Isomap [5,13], with $k = 20$ neighbors to construct the $k$-nearest-neighbors graph. To reduce noise and the influence of other motions, a third-order Butterworth low-pass filter with a cut-off frequency of 1.5 Hz is applied to the surrogate signal retrospectively.

The energy function Eq. (3) is minimized in a coarse-to-fine pyramid. This speeds up the optimization and avoids local minima. We use a downsampling factor of 0.5 and choose the number of levels such that the coarsest level has a size of $\sim 20$ pixels in each dimension. The base motions $\mathcal{V}$ and layers $\mathcal{L}$ are initialized randomly at the coarsest level. At each level, the energy function is minimized in an alternation scheme, i.e., minimization w.r.t. $\mathcal{V}$ while keeping $\mathcal{L}$ fixed, and then vice versa. This scheme is repeated 10 times on each level. A L-BFGS-B optimizer with up to 1000 iterations is used in each minimization. It is initialized with the solution of the previous alternation. The layers are constrained to be non-negative and bounded above by the image intensity maximum [12]. For non-integer positions $\boldsymbol{x}$, bilinear interpolation is used to compute intensities.

Note that $\boldsymbol{v}_l(\boldsymbol{x}, t)$ is defined in the time-dependent coordinate system of $I(\boldsymbol{x}, t)$ in Eq. (2). Intuitively, it would be preferable to model the motion of a

certain structure over time as scaled versions of a base motion by defining it in the fixed coordinate system of $L_l(\boldsymbol{x})$. However, this would require the inversion of $\boldsymbol{v}_l(\boldsymbol{x}, t)$ in each optimizer iteration to evaluate Eq. (4), which is very inefficient.

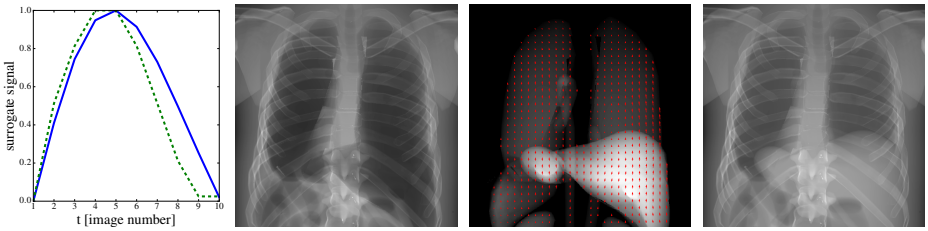
## 3 Experiments and Results

In the experiments, we evaluate the proposed method (REG-SL) on simulated and clinical X-ray sequences. The baseline method is a static layer (STAT), i.e., no motion. As alternatives, conventional 2-D/2-D registration (REG-2D) and layered motion estimation without surrogate signals (REG-L) are employed. All methods optimize the same energy Eq. (3). However, for REG-L and REG-2D, Eq. (6) is computed for each point in time and the curvature of the motion fields is regularized over time, with the weight parameter $\lambda_\tau \in \mathbb{R}$. This reduces potential bias in the evaluation, because it substitutes the inherent smoothness of the proposed surrogate-driven motion model. The parameters are empirically set to $\lambda_L = 0.05$, $\lambda_\nu = 0.025$, and $\lambda_\tau = 0.001$ such that the computed motions are visually reasonable for all methods in a pilot experiment.

### 3.1 Simulated Data

In order to densely evaluate the computed motion, simulated X-ray sequences are created by transforming two layers using known 2-D motion fields. The layers are rendered using the XCAT phantom [10] with a material-resolved renderer from CONRAD [6], where each material is assigned to a single layer. The 4-D XCAT phantom is not used directly, because then no ground truth 2-D motions would be known. The 2-D motion field of the respiratory layer is created using Eq. (2), where the base motion $\boldsymbol{\nu}_2$ is a thin-plate spline interpolation of manually annotated point motions. In the end, Gaussian and Laplacian noise with standard deviation of 1% of the image intensity range is added. The eight simulated sequences each consist of $T = 10$ images of $128 \times 128$ pixels with different layers and motions. An exemplary sequence with its constituents is shown in Figs. 1b to 1d. The ground truth motion of the respiratory layer is compared to the computed motion using the endpoint error (EE) [2]. Pixels with zero intensity in the ground-truth layer are excluded due to their unidentifiable motion, see Fig. 1c.

Table 1 shows the results of this experiment. The proposed REG-SL has the lowest EE of 2.0 mm averaged over all sequences and compensates 48% of the total motion in the images, which is represented by STAT. REG-2D and REG-L are only able to slightly decrease the EE compared to STAT. We additionally estimate the motion using the ground truth respiratory signal in the surrogate-driven motion model (REG-SL-GT), see Fig. 1a. As the EE is not substantially reduced further, the chosen method for surrogate signal extraction is validated for this application. The runtime of the methods, implemented in Python and C++, was measured on a notebook with a Core i7-3720QM processor. STAT is of course fastest, because it requires no processing. Among the registration methods, REG-2D is faster than the others with an average runtime of 33 s, because it does not need to iterate between layer and motion estimation.




(a) Surrogate signals (b) Static layer (c) Respiratory layer (d) Image

Fig. 1: Surrogate signals (true - -, estimated –), layers with overlaid motion (↑), and simulated image of XCAT experiments (best viewed in color).

Table 1: Evaluation of motion estimation methods on simulated data using runtime and endpoint error (EE) and on clinical data using TRE (mean ± std).

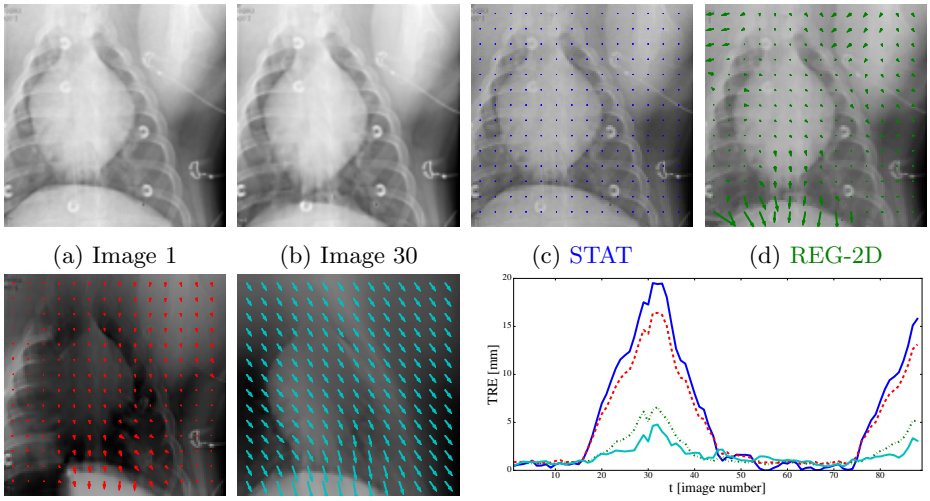
| | STAT | REG-2D | REG-L | REG-SL | REG-SL-GT |
|---|---|---|---|---|---|
| Runtime [s] | **0.0** | $33 \pm 3.5$ | $136 \pm 46$ | $48 \pm 6.8$ | $54 \pm 18$ |
| EE [mm] | $3.8 \pm 5.0$ | $3.1 \pm 3.9$ | $3.5 \pm 4.1$ | $2.0 \pm 2.5$ | $\mathbf{1.9 \pm 3.0}$ |
| TRE [mm] | $4.6 \pm 4.9$ | $3.9 \pm 5.2$ | $3.8 \pm 4.3$ | $\mathbf{2.2 \pm 3.0}$ | - |

### 3.2 Clinical Data

On clinical X-ray data, quantitative evaluation of dense 2-D motion fields or layers is challenging due to the absence of ground truth. Therefore, we resort to measuring the target registration error (TRE) at certain structures of interest. This has the drawback that the validity of the 2-D motion fields is measured only sparsely. As the target anatomy, we manually annotate structures that are known to correspond to respiratory motion, e.g., diaphragm or guidewires. The TRE is measured as the tracking error $\min_k \|C(s) - C_{\mathrm{GT}}(k)\|_2$ between each point $s$ on the computed curve $C$ and the annotated curve $C_{\mathrm{GT}}$ in mm on the detector. This experiment is performed on 6 sequences of in total 818 images with sizes of $193 - 1024$ pixels and pixel size of $0.18 - 0.43$ mm in each dimension. The images are downsampled by a multiple of two to $\sim 128$ pixels for lower runtime and memory requirements, but the error is measured in the full resolution.

The results are shown in the last row of Table 1. The total motion of the annotated structures is $4.6 \pm 4.9$ mm as represented by STAT. 2-D registration and layered motion estimation reduce the motion to 3.9 and 3.8 mm, respectively. For these methods, the extent of the reduction heavily depends on the image content. If there are few X-ray transparency effects in the region of the annotated structure, the motion is correctly estimated with REG-2D. The success of REG-L depends on the computed local minimum of the non-convex energy. REG-SL has a residual motion of $2.2 \pm 3.0$ mm, so 52% of the motion is compensated.

In the sequence of Fig. 2, our REG-SL is superior to the other methods. Transparency effects of the skin markers and the ribs deteriorate the results of REG-2D. For REG-L, neither discovered layer is anatomically plausible and thus the motions are implausible as well. In REG-SL, static structures are suppressed




(a) Image 1 (b) Image 30 (c) STAT (d) REG-2D

(e) REG-L (f) REG-SL (g) TRE over time

Fig. 2: Respiratory layer and motion (Figs. 2c to 2f) between two images (Figs. 2a and 2b) and TRE (Fig. 2g) on a real X-ray sequence (best viewed in color).

in the respiratory layer, but the diaphragm and some soft tissue is preserved. The TRE over time in Fig. 2g is most reduced by REG-SL in case of large motion.

## 4 Conclusion and Outlook

We propose a surrogate-driven motion model for layered motion estimation in X-ray fluoroscopy. The surrogate signal constrains the ill-posed optimization problem such that physiologically plausible dense 2-D motion can be estimated from X-ray images. In general, the method has little requirements. The surrogate signals can be extracted directly from the images using manifold learning, so no additional devices or synchronization are necessary. Motion estimation is independent of C-arm and table position. It can be used for thoracic or abdominal sequences, but should cover at least one breathing cycle, such that the manifold learning gives a useful respiratory signal. A restriction is the linear relationship between surrogate signal and motion. Nevertheless, its results are superior to previous approaches in our experiments. The error of 2.2 mm is in a clinically acceptable scale, e.g., for overlay navigation [4].

In future work, a more complex motion model could relax the linearity assumption, e.g., more surrogate signals per layer or a non-linear dependency between signal and motion. Another interesting point is to extend the method for estimating more than two layers. In particular, a layer with cardiac motion would be beneficial. The runtime is still too long for real-time or interactive use and should be reduced, e.g., using a GPU implementation. Furthermore, the usefulness of the computed motion must be validated for a potential clinical application, since it is only a 2-D approximation of the true 3-D motion.

*Acknowledgments.* The authors gratefully acknowledge funding of the Erlangen Graduate School in Advanced Optical Technologies (SAOT) by the German Research Foundation (DFG) in the framework of the German excellence initiative and by Siemens Healthcare. The concepts and information presented in this paper are based on research and are not commercially available.

## References

1. Auvray, V., Liénard, J., Bouthemy, P.: Multiresolution parametric estimation of transparent motions and denoising of fluoroscopic images. In: Duncan, J.S., Gerig, G. (eds.) MICCAI 2005, Part II. LNCS, vol. 3750, pp. 352–360. Springer (2005)
2. Baker, S., Scharstein, D., Lewis, J., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. Int. J. Comput. Vis. 92(1), 1–31 (2011)
3. Black, M.J., Anandan, P.: The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. Comput. Vis. Image Underst. 63(1), 75–104 (1996)
4. Brost, A., Liao, R., Strobel, N., Hornegger, J.: Respiratory motion compensation by model-based catheter tracking during EP procedures. Med. Image Anal. 14(5), 695–706 (2010)
5. Fischer, P., Pohl, T., Hornegger, J.: Real-time respiratory signal extraction from x-ray sequences using incremental manifold learning. In: ISBI. pp. 915–918. IEEE (2014)
6. Maier, A., Hofmann, H., Berger, M., Fischer, P., Schwemmer, C., Wu, H., Müller, K., Hornegger, J., Choi, J.H., Riess, C., Keil, A., Fahrig, R.: CONRAD - a software framework for cone-beam imaging in radiology. Med. Phys. 40(11), 111914 (2013)
7. Martin, J., McClelland, J., Champion, B., Hawkes, D.J.: Building surrogate-driven motion models from cone-beam CT via surrogate-correlated optical flow. In: Stoyanov, D., Collins, D., Sakuma, I., Abolmaesumi, P., Jannin, P. (eds.) IPCAI. LNCS, vol. 8498, pp. 61–67. Springer (2014)
8. McClelland, J.R., Hawkes, D.J., Schaeffter, T., King, A.P.: Respiratory motion models: a review. Med. Image Anal. 17(1), 19–42 (2013)
9. Preston, J., Rottman, C., Cheryauka, A., Anderton, L., Whitaker, R., Joshi, S.: Multi-layer deformation estimation for fluoroscopic imaging. In: Gee, J.C., Joshi, S., Pohl, K.M., Wells, W.M., Zöllei, L. (eds.) IPMI. LNCS, vol. 7917, pp. 123–134. Springer (2013)
10. Segars, W., Mahesh, M., Beck, T., Frey, E., Tsui, B.: Realistic CT simulation using the 4d XCAT phantom. Med. Phys. 35(8), 3800–3808 (2008)
11. Shizawa, M., Mase, K.: Simultaneous multiple optical flow estimation. In: ICPR. vol. 1, pp. 274–278. IEEE (1990)
12. Szeliski, R., Avidan, S., Anandan, P.: Layer extraction from multiple images containing reflections and transparency. In: CVPR. vol. 1, pp. 246–253. IEEE (2000)
13. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290(5500), 2319–2323 (2000)
14. Zhu, Y., Prummer, S., Wang, P., Chen, T., Comaniciu, D., Ostermeier, M.: Dynamic layer separation for coronary DSA and enhancement in fluoroscopic sequences. In: Yang, G.Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009, Part II. LNCS, vol. 5762, pp. 877–884. Springer (2009)
15. Zhu, Y., Tsin, Y., Sundar, H., Sauer, F.: Image-based respiratory motion compensation for fluoroscopic coronary roadmapping. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010, Part III. LNCS, vol. 6363, pp. 287–294. Springer (2010)