Predicting Slice-to-Volume Transformation in Presence of Arbitrary Subject Motion

Benjamin Hou[1], Amir Alansary[1], Steven McDonagh[1], Alice Davidson[2], Mary Rutherford[2], Jo V. Hajnal[2], Daniel Rueckert[1], Ben Glocker[1], and Bernhard Kainz[1]

[1]Biomedical Image Analysis Group, Imperial College London
[2]Division of Imaging Sciences and Biomedical Engineering, Kings College London

Abstract. This paper aims to solve a fundamental problem in intensity-based 2D/3D registration, which concerns the limited capture range and need for very good initialization of state-of-the-art image registration methods. We propose a regression approach that learns to predict rotation and translations of arbitrary 2D image slices from 3D volumes, with respect to a learned canonical atlas co-ordinate system. To this end, we utilize Convolutional Neural Networks (CNNs) to learn the highly complex regression function that maps 2D image slices into their correct position and orientation in 3D space. Our approach is attractive in challenging imaging scenarios, where significant subject motion complicates reconstruction performance of 3D volumes from 2D slice data. We extensively evaluate the effectiveness of our approach quantitatively on simulated MRI brain data with extreme random motion. We further demonstrate qualitative results on fetal MRI where our method is integrated into a full reconstruction and motion compensation pipeline. With our CNN regression approach we obtain an average prediction error of 7mm on simulated data, and convincing reconstruction quality of images of very young fetuses where previous methods fail. We further discuss applications to Computed Tomography (CT) and X-Ray projections. Our approach is a general solution to the 2D/3D initialization problem. It is computationally efficient, with prediction times per slice of a few milliseconds, making it suitable for real-time scenarios.

1 Introduction

Intensity-based registration requires a good initial alignment. General optimisation methods often cannot find a global minimum from any given starting position on the cost function. Thus, image analysis that requires registration, *e.g.*, atlas-based segmentation [2], motion-compensation [14], tracking [13], or clinical analysis of the data visualised in a standard co-ordinate system, often requires manual initalisation of the alignment. This problem gets particularity challenging for applications where the alignment is not defined by a 3D-3D rigid-body transformation. An initial rigid registration can be achieved by selecting common landmarks [3]. However, many applications, in particular motion compensation techniques, require at least approximate spatial alignment and 3D

consistency between individual 2D slices to provide a useful initialisation for subsequent automatic registration methods. Manual alignment of hundreds of slices is not feasible in practice. Landmark-based techniques can mitigate this problem, but is heavily dependent on detection accuracy and robustness of the calculated homography between locations and the descriptive power of the used landmark encoding. 2D slices also do not provide the required 3D information to establish robust landmark matching, therefore this technique cannot be used on applications such as motion compensation in fetal imaging.

Robustness of (semi-)automatic registration methods is characterised by their *capture range*, which is the maximum transformation offset from which a specific method can recover good spatial alignment. For all currently known intensity-based registration methods, the *capture range* is limited.

Contribution: We introduce a method that automatically learns slice transformation parameters relative to a canonical atlas co-ordinate system, purely from the intensity information in 2D slices. We propose a CNN regression approach that is able to predict and re-orient arbitrarily sampled slices, to provide an accurate initialisation for subsequent intensity-based registration. Our method is applicable to a number of clinical situations. In particular, we quantitatively evaluate the prediction performance with simulated 2D slice data extracted from adult 3D MRI brain and thorax phantoms. In addition, we qualitatively evaluate the approach for a full reconstruction and motion compensation pipeline for fetal MRI. Our approach can naturally be generalised to 3D-3D volumetric registration by predicting the transformation of a few selected slices. It is also applicable to projective images, which is highly valuable for X-Ray/CT registration.

Related Work: Slice-to-Volume Registration (SVR) is a key step in medical imaging, multiple 2D images can be registered together in a common world co-ordinate system to form a consistent 3D volume. This provides better visualisation for the practitioner to diagnose and/or perform operative procedures. Furthermore, it paves the way to exploit 3D medical image analysis techniques.

In literature, one can distinguish between volume-to-slice and slice-to-volume techniques. The first is concerned with aligning a volume to a given image, *e.g.*, aligning an intra-operative C-Arm X-Ray image to a pre-operative volumetric scan. This can be manually or artificially initialised and many approaches have been proposed to solve this problem. The most advanced solution, that we are aware of, uses CNNs to evaluate the spatial arrangement of landmarks automatically [13]. Other methods that can compensate for large offsets usually require use of fiducial markers [9], requiring special equipment or invasive procedures.

While our method is also applicable to the volume-to-slice problem, as shown in Exp. 3, here we focus on the slice-to-volume problem. Manual alignment of hundreds of slices to each other is much more challenging than the theoretically possible manual initialisation of volume-to-slice problems.

One target application we discuss in this paper is fetal MRI, where maternal breathing and spontaneous movement from the fetus is a major problem, requiring slice-wise re-alignment of randomly displaced anatomy [4, 8, 14, 11]. Existing methods require good initial spatial consistency between the acquired slices to

generate an approximation of the target structure, which is used for iterative refinement of SVR. Good initial 3D slice alignment is only possible through fast acquisition like single-shot Fast Spin Echo (ssFSE) and the acquisition of temporally close, intersecting stacks of slices. Redundant data covering an area of interest cannot be used from all acquired images since the displacement worsens during the course of an examination, thus redundancy has to be high and, generally, several attempts are necessary to acquire better quality data that can be motion compensated. Nevertheless, from the clinical practice, individual 2D slices are well examinable and trained experts are able to virtually realign a collection of slices mentally with respect to their real anatomical localization during diagnostics. The recent advent of deep neural network architectures [12] suggests that a learning based expert-intuition of slice transformations can also be achieved fully automatically using machine learning.

2 Method

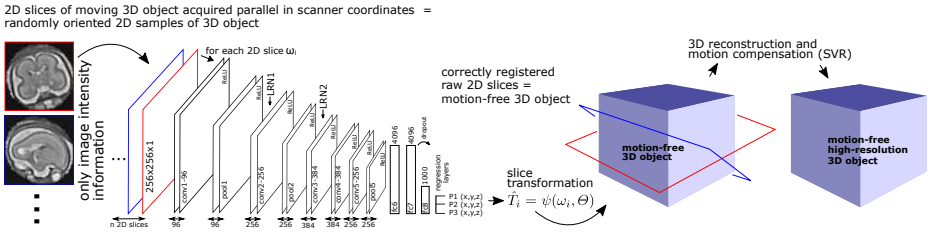


Fig. 1: Overview of Reconstruction Pipeline.

The core of our method utilises a CNN, called *SVRNet*, to regress and predict transformation parameters $\hat{T}_i$, such that $\hat{T}_i = \psi(\omega_i, \Theta)$, where Θ is the learned network parameters and $\omega_i \in \Omega$ are a series of 2D image slices that are acquired from a moving 3D object Ω. SVRNet provides a robust initialisation for intensity-based registration refinement by predicting $\hat{T}_i$ for each ω_i (see Fig. 1). We also define T_i as known ground truth parameters of ω_i during validation.

Our proposed pipeline consists of three modular components: **(I)** approximate organ localisation, **(II)** prediction of $\hat{T}_i$, and **(III)** 3D reconstruction and iterative intensity-based registration refinement.

Organ localisation, which defines a Region of Interest (ROI), can be achieved using rough manual delineation, organ focused scan sequences or automatic methods, such as [10] for example for the fetal MRI use case. For 3D Reconstruction, we use a modified SVR method [8] and initialise it with transformed ω_i using $\hat{T}_i$. Here on, we focus on the novel part of this pipeline, which is SVRNet.
Data Set Generation: ω_i, for training and validation, are generated from n motion free 3D volumes Ω. Each volume encloses a desired ROI, is centred at the origin and re-sampled to a cubic volume of length L, with spacing $1mm \times$

$1mm \times 1mm$. L/4 sampling planes, with spacing of 4mm and size $L \times L$, are evenly spaced along the Z-axis. ω_i at extremities of Ω may contain little or no content. If the variance of a particular ω_i is below a threshold of t, where $t = K \cdot \max(\sigma^2(\omega_i)), \forall i \in \Omega$ and $\sigma^2(x) = 1/N \sum_i^{N-1} |x_i - \bar{x}|^2$, then it is omitted. A higher K value will restrict ω_i to the middle portion of the volume. In our experiments, $K \approx 0.2$, which samples the central 80% of the volume.

To capture a dense permutation of $\omega_i \in \Omega_{train}$, we rotate the sampling planes about the origin whilst keeping the volume static. Ideally, all rotational permutations should be random and evenly spaced on the surface of a unit sphere. Uniform sampling of polar co-ordinates, $P(\phi, \theta)$, causes denser sampling near the poles. This can lead to an imbalance of training samples. Thus we use Fibonacci sphere sampling [5], which allows each point to represent approx. the same area. Thus sampling normals can be calculated by $P(\phi_i, \cos^{-1}(z_i))$, where $\phi_i = 2\pi i/\Phi$ and $z_i = 1 - (2i+1)/n$, $i \in 0, 1, 2, ..., n-1$. Φ is the golden ratio, as $\Phi^{-1} = \Phi - 1$, and is defined as $\Phi = (\sqrt{5}+1)/2$.

Only one hemisphere needs to be sampled due to symmetry constraints, antipode normals are the same image albeit mirrored.

Ground Truth Labels: T_i can be represented by Euler angles (six parameters: $\{r_x, r_y, r_z, t_x, t_y, t_z\}$) or Quaternions (seven parameters: $\{q_1, q_2, q_3, q_4, t_x, t_y, t_z\}$), or by defining three Cartesian anchor points within the plane (nine parameters). As Huynh et al. [6] have presented detailed analysis on distance functions for 3D rotations, we therefore implemented them as custom loss layers for regressing on rotational parameters. The loss for Euler angles can be expressed as; $\Psi_1((\alpha_1, \beta_1, \gamma_1), (\alpha_2, \beta_2, \gamma_2)) = \sqrt{d(\alpha_1, \alpha_2)^2 + d(\beta_1, \beta_2)^2 + d(\gamma_1, \gamma_2)^2}$ where $d(a, b) = \min\{|a-b|, 2\pi - |a-b|\}$, and $\alpha, \gamma \in [-\pi, \pi); \beta \in [-\pi/2, \pi/2)$. For quaternions; $\Psi_2(q_1, q_2) = \min\{\|q_1 - q_2\|, \|q_1 + q_2\|\}$, where q_1 and q_2 are unit quaternions. We have evaluated all of these options and found that the Cartesian anchor point approach yielded the highest accuracy. Hence, we use this approach in all our experiments. The anchor points can be arbitrarily selected, as long as their location remains consistent for all ω_i. In our experiments, we have chosen the centres of ω_i, p_c, and two corners p_l, p_r; where p_c = (0,0,z), $p_l = p_c$ + (-L/2,-L/2,0) and $p_r = p_c$ + (L/2,-L/2,0). To take rotation into account, each point is further multiplied by a rotation matrix R to obtain their final position in world co-ordinates. Each ω_i can thus be described by nine parameters: $p_c(x, y, z)$, $p_l(x, y, z)$ and $p_r(x, y, z)$. This approach keeps the nature of the network loss consistent as it only needs to regress in Cartesian co-ordinate space instead of a mixture of Cartesian co-ordinates and rotation parameters.

Network Design: SVRNet is derived from the CaffeNet [7] architecture. Experimentation with other architectures has revealed that this approach yields a maximum training performance whilst keeping the training effort feasible. For regression, we define multiple loss outputs; one for each p_c, p_l, p_r. SVRNet employs a multi-loss framework, which avoids over-fitting to one particular single loss [16]. Fig. 1 shows the details of the SVRNet architecture.

3D Reconstruction: As the network predicts $\hat{T}_i$ to certain degree of accuracy, we integrate an iterative intensity-based SVR motion compensation approach

to reconstruct an accurate high-resolution, motion free 3D volume, Ω, from the regression. Conventional SVR methods, *e.g.* [8], require a certain degree of correct initial 2D slice alignment in scanner co-ordinate space to estimate an initial approximation of a common volume Ω. The approximation of Ω is subsequently used as a 3D registration target for 2D/3D SVR. Our approach does not depend on good initial slice alignment and disregards slice scanner co-ordinates completely. We only use slice intensity information for SVRNet and generate an initialization for Ω using the predicted $\hat{T}_i$. We use regularized Super-Resolution and a Point-Spread-Function similar to [8] to account for different resolutions of low-resolution ω_i and high-resolution Ω. ω_i-to-Ω registration is then individually refined using cross-correlation as cost-function and gradient decent for optimization. Optimization uses three scales of a Gaussian Pyramid representation for ω_i and Ω. Robust statistics [8] identifies ω_i that have been mis-predicted and excludes them from further iterations.

3 Experiments and Results

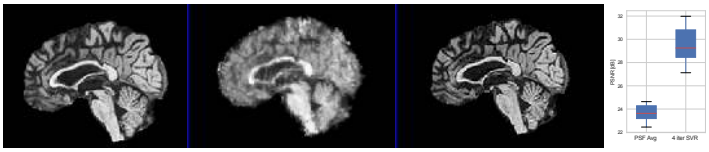
We have tested our approach on 85 randomly selected and accurately segmented healthy adult brains, on a real-world use case scenario with 34 roughly delineated fetal brain MRI scans and on 60 low-dose thorax CT scans with no organ specific segmentation. SVRNet's average prediction error for these datasets is respectively 5.6±1.07mm, 7.7±4.80mm, and 5.9±2.43mm. We evaluate 3D reconstruction performance using PSNR and average distance error in mm between ground truth locations p_c, p_l, p_r and predicted locations $\hat{p_c}, \hat{p_l}, \hat{p_r}$, such that, $e = (||p_c - \hat{p_c}|| + ||p_l - \hat{p_l}|| + ||p_r - \hat{p_r}||)/3.0$.

All experiments are conducted using the Caffe neural network library, on a computer equipped with an Intel 6700K CPU and Nvidia Titan X Pascal GPU. **Exp. 1: Segmented adult brain data** is used to evaluate our network's regression performance with known ground truth T_i. 85 brains from the ADNI data set[1] were randomly selected; 70 brains for Ω_{train} and 15 brains for $\Omega_{validation}$. Fig. 2 shows an example slice of the ground truth and the reconstructed Ω.

Each brain has been centered and re-sampled in a $256 \times 256 \times 256$ volume. Using the Fibonacci Sphere Sampling method, a density of 500 unique normals is chosen with 64 sampling planes spaced evenly apart on the Z-axis (giving a spacing of 4mm). This therefore yields a maximum of 32000 images per brain; 2.24M for the entire training set and 345K for the entire validation set. After pruning ω_i with little or no content, this figure drops to approx. 1.2M images for training and 254K for validation. Training took approx. 27hrs for 30 epochs.

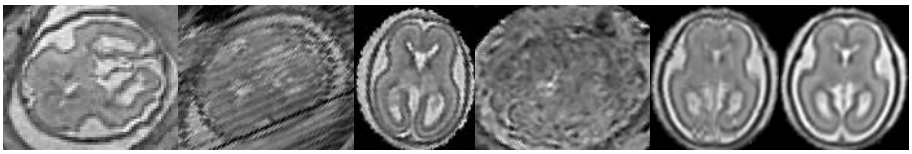
Reconstructing from $\hat{T}_i$ initialisation without SVR yields a PSNR of 23.7 ± 1.09; with subsequent SVR the PSNR increases to 29.5±2.43 when tested on 15 randomly selected test volumes after four iterations of SVR.

Exp. 2: Fetal brain data is used to test the robustness of our approach under real conditions. Fetuses younger than 30 weeks very often move a lot during examination. Fast MRI sequences allow artifact free acquisition of individual slices but motion between slices corrupts consistent 3D information. Fig. 3 shows



(a) Original (b) SVRNet (c) +SVR (d) PSNR

Fig. 2: (a): Example slice from the segmented adult brain MRI data set; (b): Reconstruction from 300 ω_i based on SVRNet regression without SVR; (c): Eight iterations of SVR. Note that SVRNet (b) predicts $\hat{T}_i$ only from image intensities without any initial world co-ordinates of the sampled slice. (d): PSNR (dB) comparing volumes of (b) and (c) to (a).



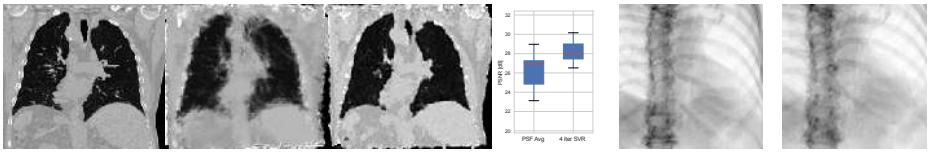
(a) ssFSE ax. (b) ssFSE sag. (c) train (d) SVR (e) SVRNet (f) +SVR

Fig. 3: (a): A single slice from a heavily motion corrupted stack of ssFSE T2 weighted fetal brain MRI; (b): Axial view of a sagittal input stack; (c): A slice at approx. the same position through a randomly selected training volume; (d): Failed reconstruction attempt using standard SVR based on three orthogonal stacks of 2D slices (the fetus moved heavily during acquisition); (e): Reconstruction based on SVRNet $\hat{T}_i$ regression; (f): Eight iterations of SVR. Note that (e) and (f) are reconstructed directly in canonical atlas co-ordinates.

that our method is able to accurately predict $\hat{T}_i$ also under these conditions. For this experiment we use ω_i from three orthogonally overlapping stacks of ssFSE slices covering the fetal brain with approx. 20-30 slices each. We are ignoring the stack transformations relative to the scanner and treat each ω_i individually. For Ω_{train}, 28 clinically approved motion compensated brain reconstructions are resampled into a $150 \times 150 \times 150$ volume with $1mm \times 1mm \times 1mm$ spacing. A density of 500 unique sampling normals has been chosen via the Fibonacci sphere sampling method with 25 sampling planes evenly spaced between -25 to +25 on the Z-axis. This gives a plane spacing of 2mm, sampling only the middle portion of the fetal brain. Training took approx. 10hrs for 30 epochs. Prediction, *i.e.*, the forward pass through the network, takes approx. 12 ms/slice.

Exp. 3: Adult thorax data: To show the versatility of our approach we also apply it to adult thorax scans. For this experiment *no organ specific* training is performed but the whole volume is used. We evaluate reconstruction performance similar to Exp. 1 and $\hat{T}_i$ prediction performance when Ω is projected on an external plane, comparable to X-Ray examination using C-Arms. The latter

provides insights about our method's performance when applied to interventional settings in contrast to motion compensation problems. 60 healthy adult thorax scans were randomly selected, 51 scans used for Ω_{train} and nine scans used for $\Omega_{validation}$. Each scan is intensity normalised and resampled in a volume of $200 \times 200 \times 200$ with spacing $1mm \times 1mm \times 1mm$. Using the Fibonacci sampling method, 25 sampling plane of size 200×200, evenly spaced between -50 and +50, were rotated over 500 normals. Training took approx. 20 hours for 60 epochs. Fig. 4c shows an example reconstruction result gaining 28dB PSNR with additional SVR. $\hat{T}_i$ prediction takes approx. 20 ms/slice for this data.



(a) original (b) SVRNet (c) +SVR (d) PSNR (e) DRR GT (f) SVRNet

Fig. 4: (a): Raw slice of low-dose thorax CT data; (b): Reconstruction based on SVRNet $\hat{T}_i$ regression; (c): Four iterations of SVR; (d): PSNR of (b) and (c) compared to (a). (e): Projection of an unseen pathological test CT volume as DRR and (e) shows a DRR at the location predicted by SVRNet when presented with the image data in (e).

We use Siddon-Jacobs ray tracing [15] to generate Digitally Reconstructed Radiographs (DRRs) from the above described data. For training, we equally sample DRRs on equidistant half-spheres around 51 CT volumes at distances of 80cm, 60cm, and 40cm, between $-90°$ and $90°$ around all three co-ordinate axes. For validation, we generate 1000 DRRs with random rotation parameters within the bounds of the training data at 60cm distance from the volumetric iso-centre. We trained on healthy volunteer data and tested on nine healthy and ten randomly selected pathological volumes (eight lung cancer and two spinal pathologies). Our approach is able to predict DRR transformations relative to the trained reference co-ordinate system with an average translation error of 106mm and 5.6° plane rotation for healthy patients, and 130mm and 7.0° average error for pathological patients. As X-Ray images are projective, the translation component is mostly degraded. A slice at 40mm may appear identical to a slightly zoomed slice at 140mm. Therefore, slice variation is based almost entirely on the orientation around the half-sphere. An example is shown in Fig. 4e,f.

Discussion & Conclusion: We have presented a method that is able to predict slice transformations relative to a canonical atlas co-ordinate system. This allows motion compensation for highly motion corrupted scans, *e.g.*, MRI scans of very young fetuses. It allows to incorporate all images that have been acquired during examination and temporal proximity is not required for good initialisation of intensity-based registration methods as it is the case in state-of-the-art methods.

We have shown that our method performs remarkably well for fetal brain data in presence of surrounding tissue and without organ specific training for low-dose thorax CT data and X-Ray to CT registration.

One limitation of our method is that SVRNet requires images to be formatted in the same way the network is trained on. This includes identical intensity ranges, spacing and translation offset removal and can be achieved with simple pre-processing methods. Furthermore, SVRNet has to be trained for a specific scenarios (*e.g.*, MRI T1, T2, X-Ray exposure, etc.). However, we show that the training region does not need to be delineated accurately and that our method is not restricted with respect to the used imaging modality and scenario.

Another limiting factor is organ symmetry, which is still an unsolved problem. This contributed the most errors throughout the experiments. ROI oversampling and automatic outlier rejection can mitigate this in real-world scenarios.

Acknowledgements: NVIDIA, Wellcome Trust/EPSRC iFIND [102431], EPSRC EP/N024494/1

References

1. Alzheimer's disease neuroimaging initiative (2017), `http://adni.loni.usc.edu`
2. Aljabar, P., et al.: Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. NeuroImage 46(3), 726 – 738 (2009)
3. Ghesu, F.C., et al.: An Artificial Agent for Anatomical Landmark Detection in Medical Images. In: MICCAI'16, Part III. pp. 229–237 (2016)
4. Gholipour, A., et al.: Robust super-resolution volume reconstruction from slice acquisitions: application to fetal brain MRI. IEEE TMI 29(10), 1739–1758 (2010)
5. González, Á.: Measurement of Areas on a Sphere Using Fibonacci and Latitude–Longitude Lattices. Mathematical Geosciences 42(1), 49 (2009)
6. Huynh, D.Q.: Metrics for 3D Rotations: Comparison and Analysis. J. Math. Imaging Vis. 35(2), 155–164 (Oct 2009)
7. Jia, Y., et al.: Caffe: Convolutional architecture for fast feature embedding. arXiv:1408.5093 (2014)
8. Kainz, B., et al.: Fast Volume Reconstruction from Motion Corrupted Stacks of 2D Slices. IEEE Trans. Med. Imag. 34(9), 1901–13 (2015)
9. Kainz, B., et al.: Fast Marker Based C-Arm Pose Estimation. In: MICCAI'08, Part II. pp. 652–659. Springer (2008)
10. Keraudren, K., et al.: Automated Localization of Fetal Organs in MRI Using Random Forests with Steerable Features. In: MICCAI'15, Part III. pp. 620–627 (2015)
11. Kim, K., et al.: Intersection Based Motion Correction of Multislice MRI for 3D in Utero Fetal Brain Image Formation. Trans. Med. Imag. 29(1), 146–158 (2010)
12. LeCun, Y., et al.: Deep learning. Nature 521(7553), 436–444 (2015)
13. Miao, S., et al.: A CNN Regression Approach for Real-Time 2D/3D Registration. IEEE Trans. Med. Imag. 35(5), 1352–1363 (2016)
14. Rousseau, F., et al.: Registration-Based Approach for Reconstruction of High-Resolution In Utero Fetal MR Brain Images. Acad Radiol 13(9), 1072 –1081 (2006)
15. Wu, J.: ITK-Based Implementation of Two-Projection 2D/3D Registration Method with an Application in Patient Setup for External Beam Radiotherapy. Insight Journal p. 784 (12 2010)
16. Xu, C., et al.: Multi-loss Regularized Deep Neural Network. IEEE Trans. Cir. and Sys. for Video Technol. 26(12), 2273–2283 (Dec 2016)