Image Reconstruction in a Manifold of Image Patches: Application to Whole-Fetus Ultrasound Imaging

Alberto Gomez[1(✉)], Veronika Zimmer[1], Nicolas Toussaint[1], Robert Wright[1], James R. Clough[1], Bishesh Khanal[1,2], Milou P. M. van Poppel[1], Emily Skelton[1], Jackie Matthews[1], and Julia A. Schnabel[1]

[1] Department of Biomedical Engineering, King's College London, London, UK
alberto.gomez@kcl.ac.uk
[2] NAAMII, Kathmandu, Nepal

Abstract. We propose an image reconstruction framework to combine a large number of overlapping image patches into a fused reconstruction of the object of interest, that is robust to inconsistencies between patches (e.g. motion artefacts) without explicitly modelling them. This is achieved through two mechanisms: first, manifold embedding, where patches are distributed on a manifold with similar patches (where similarity is defined only in the region where they overlap) closer to each other. As a result, inconsistent patches are set far apart in the manifold. Second, fusion, where a sample in the manifold is mapped back to image space, combining features from all patches in the region of the sample.

For the manifold embedding mechanism, a new method based on a Convolutional Variational Autoencoder (β-VAE) is proposed, and compared to classical manifold embedding techniques: linear (Multi Dimensional Scaling) and non-linear (Laplacian Eigenmaps). Experiments using synthetic data and on real fetal ultrasound images yield fused images of the whole fetus where, in average, β-VAE outperforms all the other methods in terms of preservation of patch information and overall image quality.

1 Introduction

Medical image reconstruction through fusion of partial captures consists of combining information from multiple images of the same object. Fusion is particularly useful when the images involved contain complementary information [7], for example when fusing Magnetic Resonance (MR) and Computed Tomography (CT) images of the brain [8] which shows more brain structures than any of the individual images, or when compounding multiple ultrasound (US) images of a

This work was supported by the Wellcome Trust IEH Award [102431]. The authors acknowledge financial support from the Department of Health via the National Institute for Health Research (NIHR) comprehensive Biomedical Research Centre award to Guy's & St Thomas' NHS Foundation Trust in partnership with King's College London and King's College Hospital NHS Foundation Trust.

F. Knoll et al. (Eds.): MLMIR 2019, LNCS 11905, pp. 226–235, 2019.
https://doi.org/10.1007/978-3-030-33843-5_21

fetus to provide whole body images [4,14]. The latter is the application targeted in this paper.

Fusion is normally a two step process: first, alignment of the images involved. Second, fusion of the aligned images. Alignment can be achieved by image registration [3]. Normally, rigid alignment is sufficient, if no motion is assumed between patches. In many cases, non-rigid motion can be expected, particularly in fetal imaging where the fetus moves frequently between acquisitions. Most research on image fusion has focused on discarding motion corrupted images or on correcting for motion using non-rigid registration [5,12]. However, registrations results are very sensitive to the registration method and the registration parameters. In the specific case of US imaging, motion correction using non-rigid registration can introduce visually abnormal patterns that degrade the quality of the reconstructed image. Moreover, the main cause of artefacts with state of the art methods is caused by motion and registration errors.

This paper introduces a novel and generic fusion framework for overlapping images (or patches) that have been aligned but may present residual registration errors and non-rigid motion artefacts. The aligned images are embedded into a manifold which separates motion corrupted patches, hence yielding a motion-free fused image without the need for non-rigid registration. The proposed method is evaluated on synthetic 2D images and on 3D fetal US.

2 Method

The key idea is illustrated in Fig. 1: we define a data set of $i = 1, \ldots, N$ image patches $I_i(\mathbf{x})$, spatially aligned (except for any non-rigid motion) and re-sampled into the same grid, so that the i-th patch only has information within a region defined by a binary mask $M_i(\mathbf{x})$. In this paper, patches are aligned rigidly using the method from [4]. Then, if patches i and j are similar in $M_i \cap M_j$, they are close neighbours in some manifold representation.

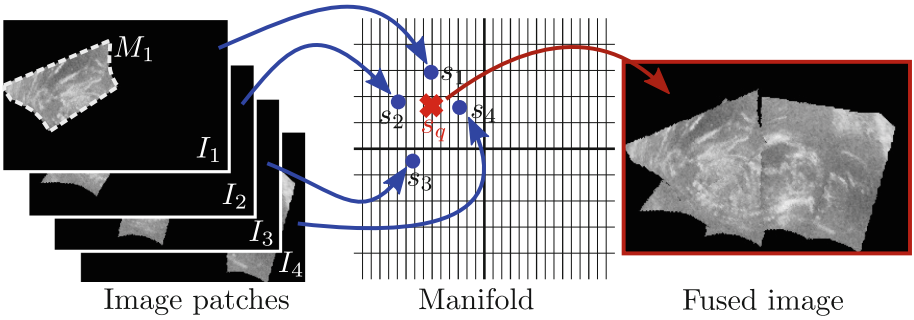

Fig. 1. Overview of the method. Patches are embedded in a manifold, and the fused image can be retrieved by mapping a sample in the manifold back to image space.

Because the corresponding location s_i of I_i in the manifold represents the entire patch, if a query location s_q from the manifold can be projected back into the image space, then the resulting image has features from nearby patches, effectively fusing the data. In this paper we compare three methods to embed the patches into a manifold: linear embedding (multi dimensional scaling, MDS [11]), non-linear embedding (Laplacian Eigenmaps, LEM [1]) and variational autoencoders (β-VAE [6]).

2.1 Image Patch Fusion with Classical Manifold Embedding

Registered images capture aligned parts of the same fetus, and differences between them are due mainly to noise, artefacts, and possibly non-corrected motion. As a result, the main variation between images can be represented in a lower-dimensional manifold. Classical manifold embedding methods work by creating a neighbourhood graph between data points (patches), which is represented as a pair-wise resemblance matrix. Then, a linear or non-linear map $\mathcal{M} : \mathbb{R}^D \longrightarrow \mathbb{R}^d$ that minimizes the change of these distances and brings the data into a $d \ll D$-dimensional space (manifold) is computed. As a matter of fact, when a linear embedding is used, the result is equivalent to a weighted average of the most similar patches using MSE as similarity criteria.

We propose to compute the pair-wise resemblance $R_{i,j}$ only in the region where the pair of patches overlap, this is $R(I_i, I_j) = r(I_i \cap M_j, I_j \cap M_i)$, where r is, here, the mean square error (MSE). This enforces that consistent patches are clustered together in the manifold, and indeed if two different patches I_i and I_j are identical in the region where they overlap, then $\mathcal{M}(I_i) = \mathcal{M}(I_j) = s \in \mathbb{R}^d$. If we could compute the inverse mapping $F = \mathcal{M}^{-1}(s)$, then F would fuse the information of I_i and I_j. However, most manifold embedding techniques are not invertible, so $\mathcal{M}^{-1}$ cannot be computed. We can estimate the fused image $\mathcal{F}(s_q)$ corresponding to a query sample s_q in the manifold by interpolating nearby samples, e.g. using Shepard's interpolation [13]:

$$\mathcal{F}(s_q) = \frac{\sum_{i \in \Omega_q} I_i M_i w_i(s_q)}{\sum_{i \in \Omega_q} M_i w_i(s_q)} \tag{1}$$

where Ω_q is a neighbourhood on the manifold around s_q and w_i is the distance-based Shepard's weight, also computed on the manifold as $w_i(s_q) = 1/\|s_q - s_i\|^2$.

2.2 Image Patch Fusion with a Variational Autoencoder

Autoencoders encode input data into a lower dimensional (latent) space, with the advantage that they provide a decoder sub-net to map the latent space back into the original space, effectively implementing the sought $F = \mathcal{M}^{-1}(s)$ mapping. β-VAEs [6,10] additionally constrain the latent space to be normally distributed, which produces consistent images from the entire manifold as the latent space is continuous by construction. As a result, β-VAEs are ideally placed to build the manifold from patch images and to retrieve a fused image from the

query manifold sample s_q. Normally, VAEs (and more broadly, neural-networks) are used to learn models of a population, and then make some predictions from unseen data. Crucially, in this work we propose to learn a model of specific subject (fetus), of which we have a large number of partial observations (overlapping patches), which yields not a population model, but a reconstruction of the subject itself. Then, instead of querying the model to obtain predictions of unseen data, the latent space (manifold) can be queried to reconstruct different poses of the subject.

We assume that each input patch $\{I_i\}_{i=1}^N$, will be similar (in the least squares sense, and within M_i) to the fused image $F_i = F(s_i)$ except for a normally distributed random noise, i.e. $I_i - F_i \cap M_i = \epsilon_i \sim \mathcal{N}(0, \sigma^2)$. As a result, $P(F|s) = P(\epsilon|s) = P(I|s)$. In consequence, the log likelihood estimator of $P(F|s)$ yields the MSE:

$$\log P(F|s) = \sum_{i=1}^{N} \log P(I_i - F_i \cap M_i|s) = C + \sum_{i=1}^{N} \frac{||I_i - F_i(s) \cap M_i||^2}{2\sigma^2} \quad (2)$$

Noting the encoder function $s = f_\phi(I)$, and adding the Kullback-Leiber (KL) divergence (weighted by β [6], which allows a trade-off between data fidelity and normal distribution of the latent space) the loss becomes:

$$\mathcal{L}(\theta, \phi, \beta; \{I\}, \{s\}) = \sum_{i=1}^{N} ||I_i - F_\theta(f_\phi(I_i)) \cap M_i||^2 + \beta KL(s_i) \quad (3)$$

The fused image can be reconstructed by sampling the latent space at $\{s_q\}$.

3 Materials and Experiments

3.1 Materials

We carry out experiments on a synthetic and real data-sets. The synthetic dataset consisted on 8 images of 128 × 128 pixels illustrating a fetus where the leg was at different locations as if captured during a kick. These images were divided into 280 overlapping patches of 40 × 40 pixels, with a 70% overlap both vertically and horizontally, to which Gaussian noise $\sim \mathcal{N}(\mu = 0, \sigma = 5)$ was added.

Experiments using 3D and 2D images from healthy fetal subjects were carried out. 3D ultrasound image sequences were acquired from two fetuses (GA 32w, 24w). Patient 1 was acquired over a head to toe sweep in which 120 volumes were acquired. Patient 2 was acquired over 5 consecutive sweeps head to toe and back, totalling 470 volumes. In both cases, data was acquired using a Philips EPIQ system with a X6-1 transducer at 4 volumes/s.

Ultrasound data was registered using the method from [4], using a grid of 1500 points distributed evenly, and each registered image was transformed and resampled into the fusion space at 1 mm^3, totalling 181×95×226 and 172×175×185 voxels per volume for each patient. Registered input volumes were sliced through

a longitudinal plane to produce 2D patches (275×184 and 352×285 pixels, respectively), which were used for the 2D experiments on real data. The β-VAE architecture was inspired by the 3D branch in [2] and represented in Fig. 2. The kernel size in all conv layers is 3×3 ($\times 3$ in 3D). Training was carried out using the Adam optimizer [9] with a learning rate of 10^{-4} and over 300 epochs.

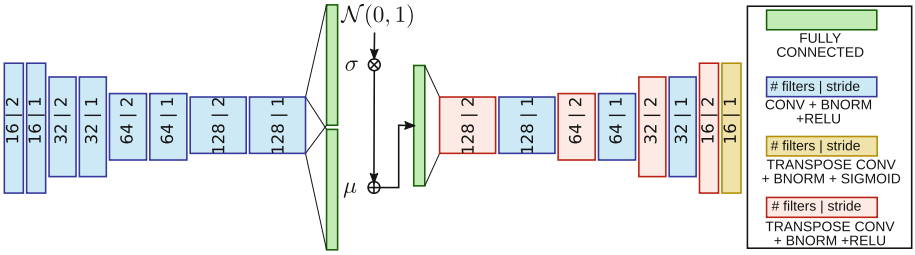

Fig. 2. Architecture of the VAE. Convolution kernels are of size 3×3 ($\times 3$ in 3D).

3.2 Experiments

Whole-body fusion of the 2D fetal ultrasound images was used for quantitative and qualitative validation. Quantitative evaluation was aimed at establishing the ability of the method to get rid of fusion artefacts (namely non-rigid motion and registration errors) while capturing the whole fetal anatomy, through the quality metric $Q(i,j) = \sqrt{\frac{1}{|M_i|}\sum_{x \in M_i} ||I_i(x) - F_j(x)||^2}$, so that $Q_{IN}(i) = Q(i,i)$, measures the RMS difference between an input patch and the same region in the fused image F_i reconstructed from the corresponding sample s_i in the manifold, therefore it measures to what extent the information in the patch was preserved. In order to measure the quality of the fusion outside the input patch I_i, we use $Q_{OUT}(i) = \frac{1}{|\Omega|}\sum_{j \in \Omega} Q(i,j)$, where the set Ω is built incorporating the patch $I_j, j \neq i$, in increasing order $Q(i,j)$, that does not intersect with patch I_i or with any of the patches already in Ω.

Further qualitative evaluation was conducted on the experiments by measuring the subjective appearance of the fused images. Three raters were presented with 500 pairs of fused images, randomly selected from a uniform sampling grid in the manifold. The raters were asked to select which image was best, or if they were of equal quality.

Both qualitative and quantitative evaluations were carried out on the synthetic and real datasets using the three methods: linear and non-linear manifold embedding and β-VAE fusion.

4 Results

The quantitative results are provided in Table 1. The pixel-wise average fusion (Avg.) was used as baseline, and compared to the two manifold embedding methods (MDS and LEM) and to the β-VAE fusion. β values in the range $[1 \cdot 10^{-4}, 1 \cdot 10^{-2}]$ were used, and a subset of this range, where results were best ($\beta \in [3 \cdot 10^{-4}, 2 \cdot 10^{-3}]$) is reported. The rows where values of β yield significantly worse results have been greyed out, so the remaining values give an idea of the range of β where results are stable. Lower values of β introduced an increasing amount of noise in the reconstruction, and higher values introduced blur, as the latent space collapsed into a single point for $\beta > 0.01$. The results show that the β-VAE approach outperforms all the other in preserving the features from individual patches. $\beta = 7 \cdot 10^{-4}$ was used for the qualitative experiments.

Table 1. Quantitative results on quality of the fused 2D image, measuring the ability of the method to preserve features from the input patches and provide a whole-fetus fusion, reported for the synthetic dataset (Synth.) and patients 1 and 2 (P1 and P2). Best values are highlighted in bold, worst are greyed out. All manifold fusion implementations outperform the naive average fusion, with the non-linear embedding (LEM) being the worst (interpolation brings the fused image outside the manifold). Overall, the β-VAE performs best, stable over a range of beta values. For patient 1, where the patch alignment is particularly good and the non-rigid motion limited over a small region covering the forearm, inter-method differences are less obvious.

		Q_{IN}			Q_{OUT}		
		Synth.	P1	P2	Synth	P1	P2
Avg		143.4 ± 33.8	153.8 ± 3.7	111.4 ± 10.6	112.9 ± 15.2	147.3 ± 3.8	96.7 ± 6.5
MDS		16.6±16.4	45.2±6.0	48.6±7.8	6.8±5.7	25.7±9.6	32.1±9.9
LEM		17.6±17.9	47.1±7.1	50.1±8.3	6.3±5.5	25.3±10.5	32.1±9.7
	β						
β-VAE	3E-4	14.1±16.4	**24.4±2.4**	17.4±2.7	6.1±3.4	**23.7±4.8**	46.6±23.5
	4E-4	14.7±17.5	25.9±2.9	**17.3±2.5**	6.4±4.0	37.0±12.5	38.5±20.2
	5E-4	13.4±15.3	25.3±3.0	17.5±2.4	6.5±3.0	24.9±5.1	44.2±23.7
	6E-4	13.6±16.5	25.6±2.7	17.8±2.9	6.1±3.6	26.4±5.3	38.8±22.1
	7E-4	**12.4±14.1**	26.4±3.1	17.8±2.8	6.2±3.1	27.9±6.7	**38.1±18.0**
	8E-4	14.2±16.3	27.2±3.1	18.0±2.9	**6.0±3.0**	30.0±6.8	40.8±21.6
	9E-4	13.6±15.1	27.7±3.1	18.2±3.0	6.8±3.8	27.5±5.0	36.4±19.5
	1E-3	13.8±15.7	28.0±3.6	18.7±3.0	6.7±4.3	29.4±6.8	39.6±19.1
	2E-3	13.0±14.3	32.5±4.2	19.6±3.3	6.3±3.6	31.6±4.9	29.7±14.4

Qualitative results in Fig. 3, show the amount of images (in %) where the fusion using the β-VAE method was judged better than the other methods, for each data-set. Overall, the β-VAE method provided better fusions with less artefacts. In the case of P1, there is no motion artefacts except for the fetal arm (c.f. second row in Fig. 4), therefore the average reconstruction is of high quality already. This explains the difference with the other data-sets.

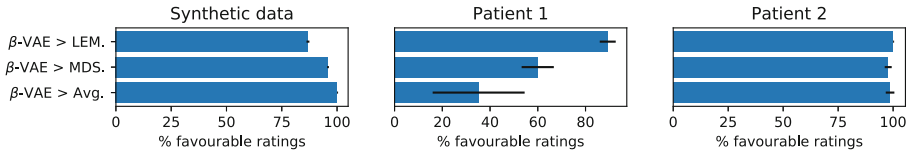

Fig. 3. β-VAE > Y indicates fraction of the times (in %) where the results with the β-VAE where considered better than with method Y by 3 raters. The bars show the average and standard deviation. The β-VAE outperforms all the other methods, with the exception of P1 where the average fusion is chosen best more often. As supported by Table 1, P1 combined good patch alignment and limited motion artefacts, so the smooth appearance of the average fusion was found to be visually best.

Examples of representative 2D fusions are shown in Fig. 4, where the fused images mapped back from the manifold sample corresponding to the patch on the left column is shown. The ability of the β-VAE to provide reconstructions without motion/blur artefacts is pointed at with white arrows. For example the second row shows, for patient 1, the β-VAE reconstructions generated from different manifold locations (corresponding to input patches marked by the red contour) that recover the entire fetus but with the arm on a different pose.

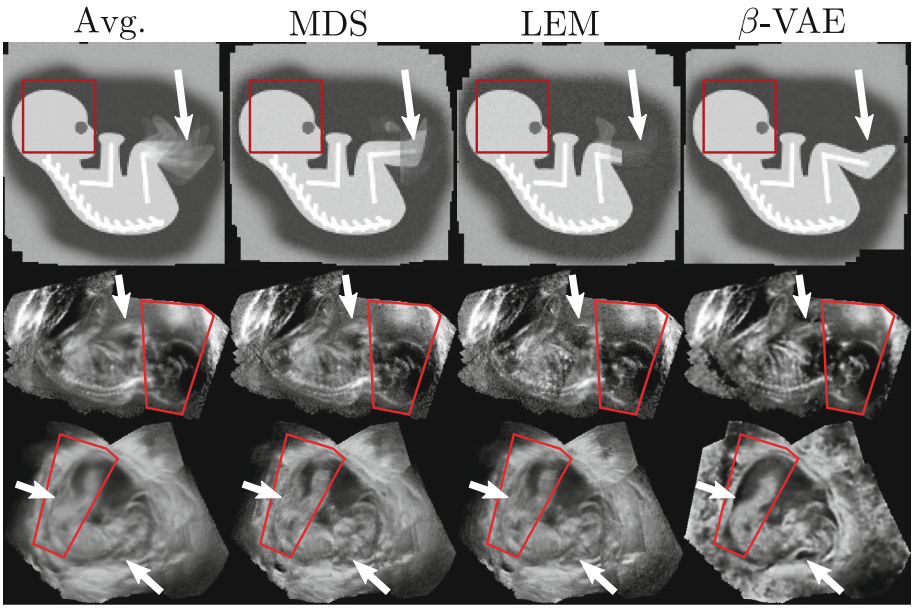

Fig. 4. 2D Whole-fetus fusions (for synthetic data, patient 1 and 2, in rows 1, 2 and 3 respectively), obtained from sampling the manifold at the location corresponding to one of the input patches. The region covered by the patch is outlined in red. White arrows indicate regions where fusion is challenging due to motion or mis-registration in input data. (Color figure online)

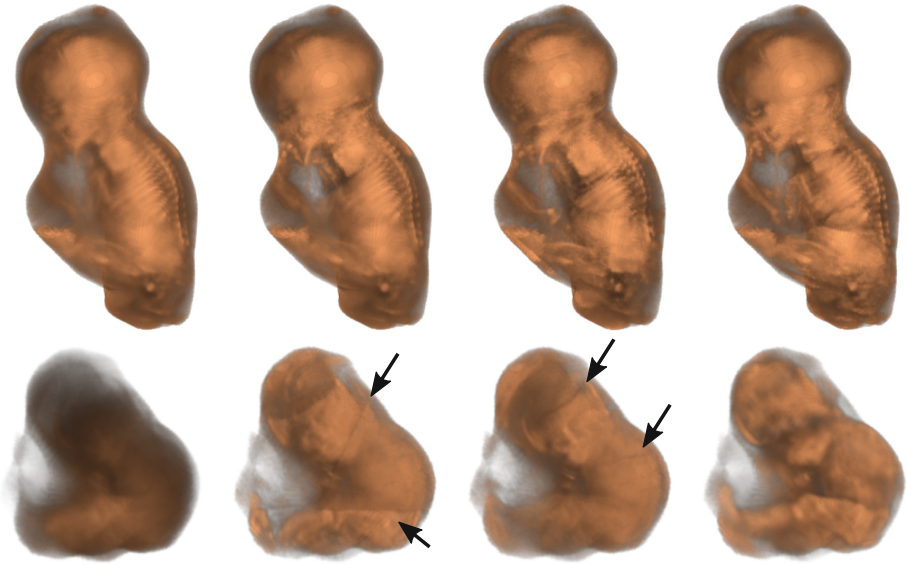

Fig. 5. Example of 3D fusion of patients 1 (top, all methods provide similar visual results) and 2 (bottom). Fusion methods, from left to right: average, MDS, LEM and β-VAE. The arrows point at some of the artefacts that were found systematically with classic manifold embedding techniques.

As a proof-of-concept, examples of the 3D version of the four methods for patients 1 and 2 are shown in Fig. 5 (as anticipated in the 2D experiments, volume renders of 3D reconstructions for patient 1 are indistinguishable). This result shows the potential of the proposed method to reconstruct high quality whole-body fetal images even from motion-corrupted input data, which would otherwise blur the result.

5 Discussion and Conclusions

We have presented a new paradigm to carry out fusion of a large amount of image patches, based on embedding the patches into a manifold through a map $\mathcal{M}$, and then sampling the manifold to reconstruct a fused image in the input image space. The proposed paradigm has been implemented using Multi-Dimensional Scaling, Laplacian Eigenmaps and a β variational autoencoder.

The inverse mapping $\mathcal{M}^{-1}$ was not available for classic manifold embedding techniques (e.g. MDS, LEM) and the fused image was obtained by Shepard's interpolation of input patches that are nearby in the manifold. Although non-linear embeddings potentially yield more accurate representation of the data, recovering the fused image through interpolation results in out-of-manifold images, which is why LEM produced worse results than the other methods. This is, to the best of our knowledge, the first time that fusion has been approached

as learning a single-instance model, where all the data are of the same object, as opposed to the common practice of creating a model that captures the variability of a population. One advantage is that this eliminates bias from under-represented cases in a training set, e.g. rare morphological abnormalities.

A limitation of the proposed framework is that it does not distinguish between inter-patch differences due to motion, noise, etc. This lends itself towards disentangled representation of these sources of variation, particularly since it may be desirable to average over noise and artefacts while separating motion. This will be investigated in future work.

The proposed method shows promising results on image fusion of rigidly pre-aligned image patches, and particularly towards a challenging task as whole body fetal image fusion. The fused images reduce the artefacts caused by non-rigid motion and misalignment by pushing the problematic patches to relatively far regions in the manifold.

References

1. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. **15**(6), 1373–1396 (2003)
2. Cerrolaza, J.J., et al.: 3D fetal skull reconstruction from 2DUS via deep conditional generative networks. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) MICCAI 2018. LNCS, vol. 11070, pp. 383–391. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00928-1_44
3. Che, C., Mathai, T.S., Galeotti, J.: Ultrasound registration: a review. Methods **115**, 128–143 (2017)
4. Gomez, A., Bhatia, K., Tharin, S., Housden, J., Toussaint, N., Schnabel, J.A.: Fast registration of 3D fetal ultrasound images using learned corresponding salient points. In: Cardoso, M.J., et al. (eds.) FIFI/OMIA -2017. LNCS, vol. 10554, pp. 33–41. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-67561-9_4
5. Heinrich, M.P., Jenkinson, M., Papież, B.W., Brady, S.M., Schnabel, J.A.: Towards realtime multimodal fusion for image-guided interventions using self-similarities. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013. LNCS, vol. 8149, pp. 187–194. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40811-3_24
6. Higgins, I., et al.: β-VAE: learning basic visual concepts with a constrained variational framework. In: ICLR, pp. 1–22 (2017)
7. James, A.P., Dasarathy, B.V.: Medical image fusion: a survey of the state of the art. Inf. Fusion **19**(1), 4–19 (2014)
8. Kavitha, C.T., Chellamuthu, C.: Multimodal medical image fusion based on integer wavelet transform and neuro-fuzzy. In: IEEE ICSIP, pp. 296–300 (2010)
9. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: ICLR (2015)
10. Kingma, D.P., Welling, M.: Auto-Encoding Variational Bayes. In: ICLR, December 2014
11. Mead, A.: Review of the development of multidimensional scaling methods. Statistician **41**(1), 27 (1992)
12. Rivaz, H., Chen, S.J.-S., Collins, D.L.: Automatic deformable MR-ultrasound registration for image-guided neurosurgery. IEEE TMI **34**(2), 366–380 (2015)

13. Shepard, D.: A two-dimensional interpolation function for irregularly-spaced data. In: Proceedings of ACM, pp. 517–524 (1968)
14. Wachinger, C., Wein, W., Navab, N.: Three-dimensional ultrasound mosaicing. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) MICCAI 2007. LNCS, vol. 4792, pp. 327–335. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-75759-7_40