Training CNNs for Image Registration from Few Samples with Model-based Data Augmentation

Hristina Uzunova, Matthias Wilms, Heinz Handels, and Jan Ehrhardt

Institute of Medical Informatics, University of Lübeck, Germany
ehrhardt@imi.uni-luebeck.de

Abstract. Convolutional neural networks (CNNs) have been successfully used for fast and accurate estimation of dense correspondences between images in computer vision applications. However, much of their success is based on the availability of large training datasets with dense ground truth correspondences, which are only rarely available in medical applications. In this paper, we, therefore, address the problem of CNNs learning from few training data for medical image registration. Our contributions are threefold: (1) We present a novel approach for learning highly expressive appearance models from few training samples, (2) we show that this approach can be used to synthesize huge amounts of realistic ground truth training data for CNN-based medical image registration, and (3) we adapt the FlowNet architecture for CNN-based optical flow estimation to the medical image registration problem. This pipeline is applied to two medical data sets with less than 40 training images. We show that CNNs learned from the proposed generative model outperform those trained on random deformations or displacement fields estimated via classical image registration.

1 Introduction

Image registration is one of the most important tasks in many medical image processing applications, e.g. for atlas-based segmentation, motion analysis or monitoring of growth processes, and therefore a variety of non-linear registration approaches have been proposed over the past three decades [15].

Inspired by the remarkable success of convolutional neural networks (CNNs) for image classification, a number of CNN-based approaches have been proposed to tackle image registration/optical flow problems in (mostly) computer vision. One line of research is to integrate CNN-based correspondence matching into registration/optical flow methods [16,11], while others successfully learned similarity metrics [14]. Recently, Dosovitskiy et al. [4] rephrased the dense optical flow problem in computer vision as a regression task, learned by CNNs in an

The final publication is available at Springer via http://dx.doi.org/10.1007/978-3-319-66182-7_26

H. Uzunova and M. Wilms contributed equally to this work.

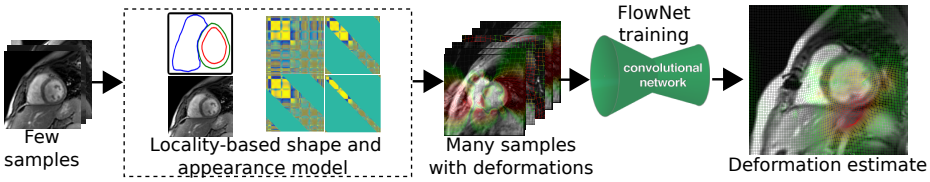

Fig. 1. Overview of the proposed model-based data augmentation approach.

end-to-end manner. Their CNN (FlowNet) is able to estimate dense deformation fields from pairs of 2D images at high frame rates and with competitive accuracy.

The success of CNNs for classification tasks heavily relies on the availability of large annotated training populations. However, for real world image registration problems, dense ground truth correspondences are rarely available and their manual generation is usually infeasible. In computer vision [4], this problem is overcome by the generation of synthetic data sets using 3D CAD models and (photorealistic) rendering. This approach is difficult to transfer to the medical field, and the lacking availability of training images of a certain kind is an even bigger challenge. This paper addresses two problems in training CNNs for image registration tasks: missing ground truth correspondences, and a small number of available training images. We aim to generate a large and diverse set of training image pairs with known correspondences from few sample images.

The usual approach to cope with few training samples is *data augmentation.* A discussion and comparison of augmentation techniques for shape modeling is given in [17]. In the context of machine learning data augmentation aims to enforce invariance of a learner to certain geometric deformations or appearance features by applying random transformations to the samples during the learning process, and, hence to improve its generalization abilities. This is a key aspect for performance improvements in recent classification and segmentation systems [12,2]. Most data augmentation schemes are manually specified, i.e. a set of geometry and intensity transformations is defined for which the task at hand is believed to be invariant, e.g. affine transformations, noise, and global changes in brightness, see e.g. [7,4]. To learn invariance related to elastic distortions, so far mostly unspecific random deformations are applied (i.e. in U-Net [12]). Only few *data-driven* augmentation techniques with transformations learned from the training data exist [10,6]. For example, in [6], non-linear transformations are learned to estimate probabilistic class-specific deformation models.

The absence of sufficiently large training populations and the unspecific data augmentation approaches currently available prevent the use of CNN-based image registration approaches like FlowNet for medical applications. We, therefore, propose a novel approach for learning representative shape and appearance models from few training samples, and embed this in a new model-based data augmentation scheme to generate huge amounts of ground truth data. Compared to [12] this allows us to synthesize more specific data and in contrast to [6] our approach also seamlessly integrates appearance related data augmentation.

The contribution of this paper is threefold: (1) A recent approach for shape modeling from few training samples [17] is extended to appearance modeling. (2) We show that this approach can be used to synthesize huge amounts of realistic ground truth training data for CNN-based medical image registration. (3) We adapt the FlowNet architecture to two medical image registration problems and show its potential to outperform state-of-the-art registration methods.

2 Methods

The training of CNNs requires huge amounts of training data, e.g. in [4] $\sim$ 22000 image pairs with dense ground truth are used to train FlowNet. Thus, the central goal of our approach is to generate many pairs of synthetic (but realistic) images $(\tilde{I}_i, \tilde{I}_j)$ with associated ground truth deformations $\phi_{i\to j}$, i.e. $\tilde{I}_j \approx \tilde{I}_i \circ \phi_{i\to j}$, from few real samples. Basically, our approach learns a statistical appearance model (SAM) [3] from the available training images and applies this model to synthesize an arbitrary number of new images with varying object shape and appearance (see Fig. 1). A common problem of classical SAMs is the limited expressiveness as the dimension of the model space is usually restricted by the number of available training images. Therefore, our appearance model adapts a recently published approach for building representative statistical shape models (SSMs) from few training data [17]. This allows us to generate highly flexible SAMs from few real samples. We begin by briefly describing statistical appearance models (SAMs), followed by our adaption of the approach presented in [17].

2.1 Statistical appearance models

Given are a set of n training images $I_1, \dots, I_n$; $I_i : \Omega \to \mathbb{R}$, $\Omega \subset \mathbb{R}^2$, and for each image I_i a set of m landmarks $\boldsymbol{s}_i = [s_{i,1}, \dots, s_{i,m}]^T \in \mathbb{R}^{2m}$ with $s_{i,r} = [x_{i,r}, y_{i,r}]^T$. These landmarks describe the shape of the object(s) of interest and are assumed to be in correspondence across the population and normalized using Procrustes analysis [3]. To generate the shape model from the shape vectors $\boldsymbol{s}_i$, the mean shape $\boldsymbol{s}_0$ and the orthonormal shape basis $\mathbf{P_S} = (\boldsymbol{p}_1 | \dots | \boldsymbol{p}_k)$ given by the first $k < n$ eigenvectors of the data covariance matrix $\mathbf{C_S}$ are calculated. New shapes can now be generated using normally distributed shape parameters $w_j \sim N(0, \lambda_j)$ with the variance λ_j equal to the corresponding eigenvalue:

$$\hat{\boldsymbol{s}} = \boldsymbol{s}_0 + \sum_{j=1}^{k} w_j \boldsymbol{p}_j \ . \tag{1}$$

The appearance model of a SAM is defined with respect to the mean shape $\boldsymbol{s}_0$, i.e. each training image I_i is *shape normalized* by warping the shape vector $\boldsymbol{s}_i$ to $\boldsymbol{s}_0$. We use a multi-level B-spline scattered data approximation [8] to define the non-linear warp φ_i and choose a number of levels that fulfill $\max_r \|s_{0,r} - \varphi_i(s_{i,r})\| < \epsilon$. In our experiments this approach leads to visually more realistic deformations compared to thin-plate-splines [3] or piecewise-affine warps [9].

The appearance covariance matrix $\mathbf{C_A}$ is computed from the shape normalized images $I_i \circ \varphi_i$ sampled at positions $\boldsymbol{x}_j \in \Omega_0$. A PCA results in a mean image I_0 and eigenimages $\mathbf{P_A} = (A_1|\dots|A_\kappa)$ defining the appearance model $\hat{I} = I_0 + \sum_{j=1}^{\kappa} \gamma_j A_j$. Again, the appearance parameters γ_j are assumed to be normally distributed and we can generate new image instances by (1) sampling shape parameters to define the shape $\hat{\boldsymbol{s}}$ and calculating the inverse warping function $\hat{\varphi}^{-1}$, (2) sampling appearance parameters to generate $\hat{I}$, and (3) warping the image $\tilde{I} = \hat{I} \circ \hat{\varphi}^{-1}$. However, SAMs strongly suffer from the high-dimension-low-sample-size (HDLSS) problem because the dimension of the embedding space is high ($\sim$ number of pixels and landmarks) compared to the number of training images. This results in a limited generalization ability and thus hampers their applicability in the intended deep learning scenario.

2.2 Locality-based statistical shape and appearance models

Recently, a new approach to tackle the HDLSS problem of SSMs was proposed [17]. This locality-based approach assumes that local shape variations have limited effects in distant areas. To measure the distance $dist(\cdot,\cdot)$ between landmarks simple euclidean or geodesic contour distances can be used, but more elaborate distances incorporating prior knowledge and multi-object scenarios are also possible (see [17]). To enforce the locality assumption during model generation a distance threshold τ is defined and the correlation of distant landmark positions $\bar{s}_i$,$\bar{s}_j$ of the mean shape is set to zero:

$$\boldsymbol{R}_\tau = \{\rho\}_{i,j} \quad \text{with } \rho_{i,j} = \begin{cases} \frac{\text{cov}(\bar{s}_i,\bar{s}_j)}{\sigma_i \sigma_j}, & \text{if } dist(\bar{s}_i,\bar{s}_j) < \tau \\ 0, & \text{else} \end{cases} . \tag{2}$$

Here, $\boldsymbol{R}_\tau$ denotes a correlation matrix related to the modified covariance matrix $\boldsymbol{C}_\tau = (diag(\boldsymbol{C}))^{1/2} \boldsymbol{R}_\tau (diag(\boldsymbol{C}))^{1/2}$. Finally, the eigenvectors of $\boldsymbol{C}_\tau$ form a new shape basis $\mathbf{P}_\tau$. For small thresholds τ, each eigenvector tends to reflect only local shape variations present in the training set, and because $rank(\hat{\boldsymbol{C}}_\tau) \gg rank(\boldsymbol{C})$ now a large number $k > n$ of eigenvectors can be selected for shape modeling in Eq.(1). The manipulation of the correlation matrix (instead of directly changing the covariances) will preserve the total variability in the training set.

By selecting a set of thresholds $\tau_1 > \tau_2 > \dots > \tau_l$, a single multi-level shape model can be build that incorporates shape variations at different levels of locality. Let $span(\mathbf{P}_{\tau_1}) = \mathcal{P}_1 \in \mathcal{G}(k_1, 2m)$ and $span(\mathbf{P}_{\tau_2}) = \mathcal{P}_2 \in \mathcal{G}(k_2, 2m)$ the subspaces of two locality-models ($\mathcal{G}(k_i, 2m)$ denotes a Grassmann manifolds) the k_2-dimensional subspace nearest to $\mathcal{P}_2$ containing $\mathcal{P}_1$ is sought ($k_2 \geq k_1$):

$$\mathcal{P}_{1+2} = \arg \min_{\mathcal{P} \in \mathcal{G}(k_2,2m)} d_{\mathcal{G}(k_2,2m)}(\mathcal{P}, \mathcal{P}_2) \quad \text{subject to } \mathcal{P}_1 \subseteq \mathcal{P} , \tag{3}$$

Here, $d_{\mathcal{G}(k_2,2m)}(\cdot,\cdot)$ denotes a geodesic distance between subspaces. The basis vectors of $\mathcal{P}_{1+2}$ and the associated eigenvalues can be efficiently computed as shown in [17]. By successively solving Eq.(3) for the remaining levels of locality

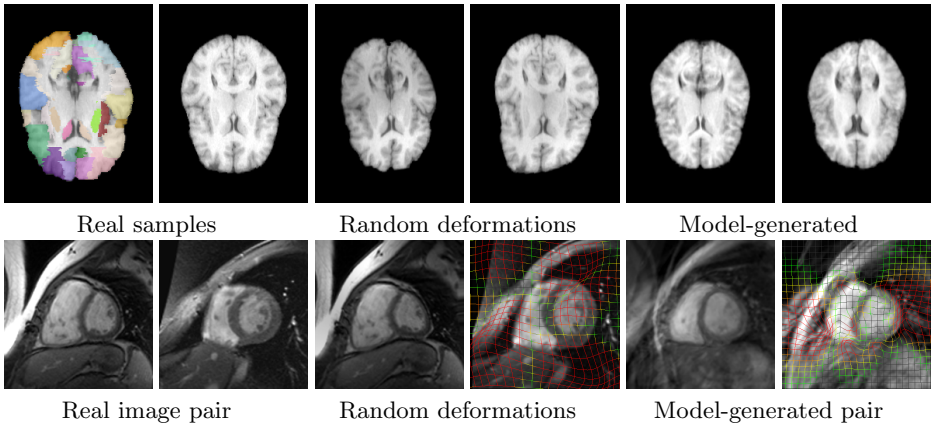

Fig. 2. Exemplary illustration of both data sets and the data augmentation approaches. Top row: LBPA40 brain data with ground truth labels (1st image). Bottom row: Image pairs of the cardiac MRI data with overlayed deformations generated by the data augmentation approaches (random deformations and the novel model-based approach).

$\tau_3, \ldots, \tau_l$, a subspace $\mathcal{P}_{1+2+\ldots+l}$, which includes global as well as very local shape variations is found (see [17] for details).

In [17], this locality-based approach is only defined for SSMs. Here, we extend it to appearance models, by using the Euclidean distance between sampling positions $\boldsymbol{x}_j \in \Omega_0$ in the image plane and associated threshold $\vartheta_1 > \vartheta_2 > \ldots$ to enforce uncorrelated image intensities in Eq.(2). To define the thresholds for multiple resolution levels, we found $\vartheta_1 = \max_{i,j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$, $\vartheta_i = \frac{\vartheta_{i-1}}{2}$ to be a reasonable choice where the number of levels depends on the required locality.

2.3 Model-based data augmentation for learning image registration

The locality-based shape and appearance model defined in Sec. 2.2 elegantly combines global and local variabilities in a single model, described by shape vectors, eigenimages, and the associated eigenvalues. Assuming Gaussian distributions for the shape and appearance parameters, we can directly apply the method described in Sec. 2.1 to generate new random images. The shape vectors $\hat{\boldsymbol{s}}_i$ and $\hat{\boldsymbol{s}}_j$ associated with the random samples $\tilde{I}_i$ and $\tilde{I}_j$ are used to compute the dense deformation $\phi_{i \to j}$ by a multi-level B-spline approximation [8], see Sec. 2.1. Clearly, the accuracy of this deformation decreases with increasing distance from landmarks, which will be discussed in Sec. 3.

3 Experiments and Results

Data. To our knowledge, dense 3D registration with CNNs is currently computationally infeasible, and we, therefore, use two 2D inter-patient registration problems for our evaluation.

Brain MRI: We extract corresponding transversal slices from affinely pre-registered image volumes of 40 patients of the LPBA40 data set [13]; see Fig. 2 for examples. For each 2D image 100 landmarks on the brain contour and 12 inner brain landmarks are defined for shape modeling. The average Jaccard overlap of 20 brain structures is used to assess the registration accuracy.

Cardiac MRI: We extract end-diastolic mid-ventricular short-axis slices from 32 cine MRI images [1]. Shape correspondences are defined by 104 landmarks located on left ventricle (LV) epicardium, and right+left ventricle endocardium. For the evaluation we compute average symmetric contour distances for the RV+LV endocard and LV epicard contours.

Experimental setup. There are only few approaches for CNN-based end-to-end training for dense image registration, and currently, FlowNet [4] is the best known among these. Therefore, the pre-trained FlowNet-S is used as starting point for all CNN experiments, followed by a fine-tuning with ground truth image pairs generated as detailed below. We adapted the data augmentation steps included in the FlowNet architecture to fit our image data (e.g. by removing color manipulations) [1]. The general goal of the 3 experiments conducted, is to investigate our initial hypotheses that (1) fast CNN-based registration can achieve competitive accuracy on medical data given sufficient training data, and that (2) the proposed data-driven, model-based augmentation approach outperforms available generic, but highly unspecific methods.

FlowNet-Reg: In this experiment, we define ground truth deformation fields by an affine landmark-based pre-registration followed by a diffeomorphic variational non-linear registration of all training image pairs. Pairwise registration will result in $n(n-1)$ image pairs, which might be not sufficient for training if n is small. The chosen registration method is freely available in the ITK framework and among the best performing methods on LPBA40 (see [5] for parameters).

FlowNet-Random: Dense smooth random deformations as suggested in [12] are applied to all training images, and combined with smooth local brightness changes. With this approach, an arbitrary number of image pairs with known ground truth can be generated, but both images of each pair are deformed versions of the same input image (see Fig. 2) and the deformations are unspecific.

FlowNet-SAM: The proposed locality-based shape and appearance model (see Sec. 2.2) is applied to generate image pairs and corresponding ground truth deformation as detailed in Sec. 2.3. The multi-object distance suggested in [17] with 4 (Brain)/3 (Cardiac) levels of locality is used for SSM generation.

The accuracy of the multi-level B-spline deformations used to infer dense displacement fields from landmark correspondences in Sec. 2.1 decreases far away from landmarks, and this results in a blurred appearance model in these regions as visible in Fig. 2. One solution is to spread landmarks over the whole image region, however, this is impractical in many applications. Instead, we adapt FlowNet and use a weighted loss function during training, with weights of 1 inside the objects (e.g. heart) that decrease to 0 far away from the contour.

[1] Architecture and trained weights: http://imi.uni-luebeck.de/node/1019

Table 1. Results of the experiments on both data sets. Given are mean Jaccard coefficients (Brains)/ contour distances in *mm* (Cardiac) over 5-folds with respect to the ground truth segmentations/landmarks. Shown is FlowNet trained on 4 data sets. Pre-trained: trained on synthetic chair data (see [4]); Reg: fine-tuned on VarReg (training data). Random: random deformations; SAM: data augmentation using the proposed models. Note the different number of training samples (2nd, 4th column). Superscripts indicate statistically significant differences to FlowNet SAM ($\diamond$:$p < 0.01$, $\star$:$p < 0.001$).

	Brains (Jaccard)		**Cardiac** (contour dist.)	
Method	**# train**	**mean(±std)**	**# train**	**mean(±std)**
Before Reg		$0.460 \pm 0.063^{\star}$		$6.163 \pm 2.472^{\star}$
VarReg (training data)		0.563 ± 0.053		2.250 ± 0.755
VarReg (test data)		$0.562 \pm 0.051^{\diamond}$		$3.437 \pm 2.427^{\star}$
FlowNet (pre-trained)	22232	$0.507 \pm 0.053^{\star}$	22232	$8.171 \pm 6.981^{\star}$
FlowNet-Reg	945	$0.547 \pm 0.049^{\star}$	600	$3.053 \pm 0.910^{\star}$
FlowNet-Random	9698	$0.505 \pm 0.077^{\star}$	9698	$7.785 \pm 5.430^{\star}$
FlowNet-SAM	9572	0.568 ± 0.042	9572	2.670 ± 0.930

Results. A 5-fold cross-validation is applied for all experiments on both image data sets. To compute a baseline accuracy, variational registration is applied to the test data without any landmark information for the brain images and using heart ROI masks for the cardiac data. Note that cardiac inter-patient registration is very challenging for intensity-based registration methods due to the large anatomical variations between patients (see Fig. 2). The results are summarized in Tab. 1 and show that FlowNet trained with model-generated data (*FlowNet-SAM*) outperforms all other methods with high significance (paired t-test, $p < 0.001$, except for brain images $p < 0.01$). The registration of one image pair (256×256) needs $0.05s$ on the GPU. *FlowNet-Random* and *FlowNet-SAM* were trained with ca. 10000 samples, which in our experiments was found to be a lower bound. The Jaccard coefficients for the brain scenario of the registration method and *FlowNet-SAM* are comparable to the 3D values of state-of-the-art methods [5]. Interestingly, for the difficult cardiac registration problem (see VarReg results), pre-trained FlowNet fails, which might suggest that the filters learned on the synthetic chair data (see [4]) are useless in this scenario. Fine-tuning with the proposed approach, however, greatly improves the results. As assumed, fine-tuning with random deformations does not provide much meaningful information for medical data, resulting in poor registration accuracy.

4 Discussion and Conclusion

In this work, we propose the use of CNN-based image registration for medical image data and present a novel model-based data augmentation scheme to allow for deep learning on small training populations. The results of our evaluation confirm our initial hypotheses that CNN-based registration can achieve competitive accuracy on medical data and that the proposed model-based augmentation

approach outperforms unspecific augmentation schemes. We can furthermore show that simple but specific fine-tuning of the FlowNet architecture designed and pre-trained for/with completely different data gives surprisingly good results. We, therefore, strongly believe that CNN-based image registration has the potential to outperform state-of-the-art medical image registration methods in the future. Currently, FlowNet is limited to 2D registration problems. However, this limitation does not apply to the proposed data augmentation approach, which readily generalizes to 3D.

References

1. Andreopoulos, A., Tsotsos, J.K.: Efficient and generalizable statistical models of shape and appearance for analysis of cardiac MRI. Med Image Anal 12(3), 335–357 (2008)
2. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: Learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Wells, W., Sabuncu, M., Unal, G., Joskowicz, L. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer (2016)
3. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Trans PAMI 23(6), 681–685 (2001)
4. Dosovitskiy, A., Fischer, P., Ilg, E., et al.: Flownet: Learning optical flow with convolutional networks. In: CVPR 2015. pp. 2758–2766 (2015)
5. Ehrhardt, J., Schmidt-Richberg, A., Werner, R., Handels, H.: Variational registration: A flexible open-source itk toolbox for nonrigid image registration. In: Handels, H., Deserno, T.M., Meinzer, H.P., Tolxdorff, T. (eds.) Bildverarbeitung für die Medizin 2015. pp. 209–214. Springer (2015)
6. Hauberg, S., Freifeld, O., Larsen, A.B.L., et al.: Dreaming more data: Class-dependent distributions over diffeomorphisms for learned data augmentation. In: AISTATS 2016. pp. 342–350 (2016)
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS 2012. pp. 1097–1105 (2012)
8. Lee, S., Wolberg, G., Shin, S.Y.: Scattered data interpolation with multilevel b-splines. IEEE Trans Vis Comput Graph 3(3), 228–244 (1997)
9. Matthews, I., Baker, S.: Active appearance models revisited. IJCV 60(2), 135–164 (2004)
10. Miller, E.G., Matsakis, N.E., Viola, P.A.: Learning from one example through shared densities on transforms. In: CVPR 2000. pp. 464–471 (2000)
11. Revaud, J., Weinzaepfel, P., Harchaoui, Z., Schmid, C.: Epicflow: Edge-preserving interpolation of correspondences for optical flow. In: CVPR 2015. pp. 1164–1172 (2015)
12. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer (2015)
13. Shattuck, D.W., Mirza, M., Adisetiyo, V., et al.: Construction of a 3d probabilistic atlas of human cortical structures. Neuroimage 39(3), 1064–1080 (2008)
14. Simonovsky, M., Gutiérrez-Becker, B., Diana, M., Navab, N., Komodakis, N.: A deep metric for multimodal registration. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9902, pp. 10–18. Springer (2016)

15. Sotiras, A., Davatzikos, C., Paragios, N.: Deformable medical image registration: A survey. IEEE Trans Medical Imaging 32(7), 1153–1190 (2013)
16. Weinzaepfel, P., Revaud, J., Harchaoui, Z., Schmid, C.: Deepflow: Large displacement optical flow with deep matching. In: CVPR 2013. pp. 1385–1392 (2013)
17. Wilms, M., Handels, H., Ehrhardt, J.: Multi-resolution multi-object statistical shape models based on the locality assumption. Med Image Anal 38, 17–29 (2017)