Semantically Guided 3D Abdominal Image Registration with Deep Pyramid Feature Learning

Mona Schumacher[1,2], Daniela Frey[3], In Young Ha[1], Ragnar Bade[2], Andreas Genz[2], Mattias Heinrich[1]

[1]Institute of Medical Informatics, University of Lübeck
[2]MeVis Medical Solutions AG, Bremen
[3]University of Applied Sciences, Lübeck
Mona.Schumacher@mevis.de

Abstract. Deformable image registration of images with large deformations is still a challenging task. Currently available deep learning methods exceed classical non-learning-based methods primarily in terms of lower computational time. However, these convolutional networks face difficulties when applied to scans with large deformations. We present a semantically guided registration network with deep pyramid feature learning that enables large deformations by transferring features from the images to be registered to the registration networks. Both network parts have U-Net architectures. The networks are trained end-to-end and evaluated with two datasets, both containing contrast enhanced liver CT images and ground truth liver segmentations. We compared our method against one classical and two deep learning methods. Our experimental validation shows that our proposed method enables large deformation and achieves the highest Dice score and the smallest surface distance of the liver in constrast to other deep learning methods.

1 Introduction

Medical imaging techniques are important for diagnosis and treatment planning. In order to be able to ensure a detailed analysis of the structures to be examined, different images with complementary information are acquired and the information of these images have to be combined in order to be able to make a valid decision. One important example is the registration of contrast enhanced liver CT images. The organs in the abdomen are, due to their soft tissue structures, flexible and can be deformed by the (respiratory) movement of the patient, which makes a mapping of the liver and vascular systems necessary, e.g., to plan a liver surgery.

To address this issue, most classical image registration algorithms iteratively optimise a similarity metric in order to spatially align images, which leads to long computation times [1].

Convolutional neural networks overcome the drawback of a long processing time since the iterative optimization takes place during training and not during inference [2,3]. Most of the currently available approaches show a good performance on regions with small deformations, e.g. brain images, but have difficulties registering images with large deformations.

The aim of this work is to overcome the drawback using learning-based networks and enable large deformations. Since it is rather difficult to create ground truth data for a registration, most methods are unsupervised. One example is VoxelMorph that was introduced by Balakrishnan et al. in 2018 [2]. It is based on the intensity values in the image and can use mutual information as similarity metric to maximize the image correspondence. Besides approaches that learn in an unsupervised manner, there are also methods that use other expert information as guidance. Ha et al. in 2020 [3] focus on the 2D registration of segmentations as regions of interest and combine a U-Net segmentation and a registration network that are jointly optimized. In 2D optical flow learning, the concept of feature pyramids proposed in the PWC-Net has led to huge advances [4]. Our work combines and extends the ideas of Sun et al. and Ha et al. First, we adapted the networks to enable 3D registrations. Additionally, we overcome the drawback of the registration network, which outputs a coarse deformation field. Instead we use a U-Net architecture that additionally has a decoder path and outputs a larger deformation field. To further guide the registration, we transfer the information of each resolution of the decoder path of the semantic network to the corresponding resolution of the encoder path of registration network with skip connections.

2 Material and Method

A general overview of our proposed method is shown in Fig. 1. Our method consists of two parts: a feature extraction and a registration part. The first part extracts the semantic features by generating a liver segmentation of the fixed and the moving image with two U-Nets that have shared weights. The deformation field is estimated in the second part by two registration networks that have U-Net architecture, hence an encoder and a decoder path. The input of the first registration network are the concatenated outputs of the fixed and moving feature extraction networks. In the downsampling path, the feature maps of the feature extraction networks are concatenated to the corresponding resolution. The output of the first registration network is an initial deformation field. The input of the second registration network are the outputs of the feature extraction networks that are warped with the initial deformation field. The feature maps of the upsampling path of the feature extraction network are also warped with the initial deformation field and are concatenated to the corresponding resolution level in the encoder path. The resulting deformation field of the second registration network is added to the initial deformation field. This sequential generation of the final deformation field enables us to align large deformations.

2.1 Semantic Feature Extraction

To extract semantic features in the images to be registered, two U-Nets with shared weights are used to generate liver probabilities of the fixed and the moving image. The U-Net contains three downsampling steps with convolutions with a kernel size of $3 \times 3 \times 3$, stride of 2 and a padding of 1, followed by 3D instance normalization and leaky ReLU with a slope of 0.1. The decoder part is structured parallel to the encoder part but contains only two upsampling steps. Additionally, the network contains skip connections from the encoder path to the decoder path. The output corresponds to the halved resolution of the input images. The semantic loss is a weighted cross-entropy with weights computed by the square root of the inverse class frequency for each label.

2.2 Registration Networks

The deformation field is estimated by two registration networks with the same architecture. It comprises three downsampling and upsampling steps. In contrast,

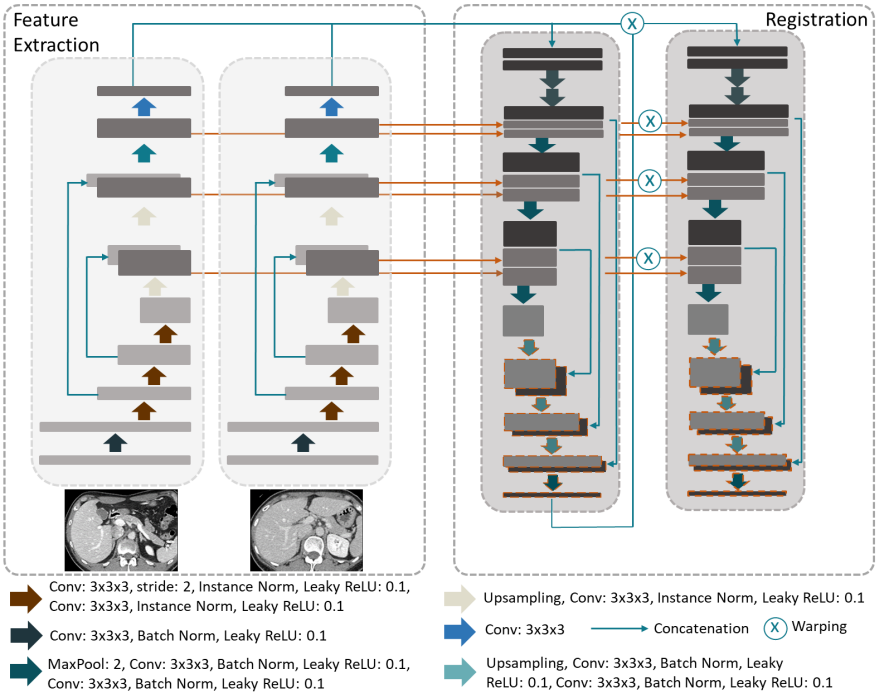

Fig. 1. Pipeline of semantically guided registration. The feature extraction part consists of two U-Nets with shared weights that extract the features of the moving and the fixed image. These features are transferred to the second part, the registration, which consists of two U-Nets connected in series. The first U-Net creates an initial deformation field, which is finalized by the second U-Net. All added elements (compared to the original architecture) are marked orange.

the original approach consists of four downsampling steps without upsampling. The downsampling steps contain convolutions with a kernel size of $3 \times 3 \times 3$, stride of 2 and padding of 1, followed by 3D batch normalization and leaky ReLU with a parameter of 0.1. The decoder part is structured parallel to the encoder part. The deformation and regularization loss is calculated as described in [3].

2.3 Image Data

To train, validate and test our approach, and to compare it to other registration techniques, two contrast enhanced liver CT datasets are used.

The first dataset consists of 170 images: 85 images of the venous phase and 85 images of the late-venous phase. For all of these images, ground truth labels of the liver are available. The data has been manually segmented and reviewed by three experienced radiologic technicians. The CT scans and the corresponding segmentations are resampled to a voxel size of $0.9 \times 0.9 \times 1.3$ mm^3 and an initial affine registration is applied. Around the center of the segmented liver, the image is cropped to $304 \times 240 \times 144$ voxels and zero padded if necessary. The segmentations of the liver of the corresponding phases are processed in the same way. 43 images of each of the two phases are assigned to the training data set, 42 image of each phase to the test data set.

The second dataset is a publicly available dataset from the Learn2Reg MICCAI Registration Challenge 2020[1] that consists of 30 abdominal CT scans of different patients of the portal venous phase. The dataset includes ground segmentations of 13 organs, in particular also of the liver. In contrast to the first dataset, the initial average liver Dice score is much lower so that even larger deformations are required. The images are already preprocessed to the same voxelsize of $2.0 \times 2.0 \times 2.0$ mm^3, a spatial dimension of $192 \times 160 \times 256$ voxels and affine registered. For the following experiments only the liver segmentation is taken into account. 20 images are included in the training process, while 10 are used as test dataset.

3 Results

All networks are trained across patients to ensure a sufficient number of image pairs and obtain a good generalization. During training, two random images are chosen as fixed and moving images so that the choice of phases is randomized. We have performed an ablation study for the proposed pipeline. First, we adapted the original approach of Ha et al. [3] to a 3D approach and trained it with the two datasets. The resulting deformation field is of the size $9 \times 7 \times 4$ voxels for our dataset and $6 \times 5 \times 8$ for the Learn2Reg dataset. The deformation field is upsampled to the cropped image size. The networks are trained end-to-end with an Adam optimizer with a learning rate of 0.01. A number of 300 epochs is carried out for the datasets.

[1] https://learn2reg.grand-challenge.org/Datasets/ Task 3

Table 1. Results of deformable registration with our test dataset (upper part) and the Learn2Reg test dataset (lower part). Average Dice Scores, mean of the Average Surface Distance (ASD), 95% Hausdorff Distance (HD95), lowest 30% of the Dice coefficients (Dice30), standard deviation of the Jacobian determinant, and the percentage of negative elements of the Jacobian are presented.

Method	Dice	ASD	HD95	Dice30	JacStd	JacDet < 0
Affine	0.71 ± 0.12	11.74 ± 5.68	34.79 ± 17.56	0.58 ± 0.09	-	-
deeds	0.82 ± 0.11	7.73 ± 5.54	29.23 ± 20.44	0.70 ± 0.08	0.42	0.0
VoxelMorph	0.74 ± 0.11	10.56 ± 5.42	33.20 ± 17.25	0.63 ± 0.09	0.36	0.15
Ha et al.	0.87 ± 0.11	4.61 ± 2.02	16.24 ± 9.77	0.83 ± 0.04	0.77	4.62
Ours	0.90 ± 0.03	4.07 ± 2.30	15.76 ± 10.75	0.87 ± 0.02	1.55	6.22
Ours+Skip	0.91 ± 0.03	3.90 ± 2.06	15.36 ± 10.21	0.87 ± 0.02	1.30	6.12
Affine	0.62 ± 0.10	15.39 ± 4.60	49.27 ± 13.21	0.51 ± 0.06	-	-
deeds	0.83 ± 0.06	7.72 ± 4.03	35.54 ± 17.45	0.77 ± 0.06	0.49	3.78
VoxelMorph	0.68 ± 0.08	12.88 ± 3.99	45.04 ± 12.85	0.59 ± 0.05	0.35	0.38
Ha et al.	0.72 ± 0.07	10.86 ± 3.46	35.73 ± 11.77	0.65 ± 0.04	0.25	0.19
Ours	0.77 ± 0.08	9.16 ± 3.80	33.22 ± 14.61	0.67 ± 0.04	0.46	2.01
Ours+Skip	0.79 ± 0.07	8.56 ± 3.48	31.78 ± 13.80	0.71 ± 0.04	0.46	1.75

Next, we adapted the registration networks, whereas the learning parameters remain as in the first experiment. First, the decoding path is added. Additionally, the number of downsampling steps is reduced to three. The downsampling path therefore results in a larger feature map of dimension $19 \times 15 \times 9$ for our dataset and $12 \times 10 \times 16$ for the Learn2Reg dataset. The final deformation field has the size of the input feature map. Second, further skip connections are introduced to additionally guide the registration network with semantic features as shown in Fig. 1.

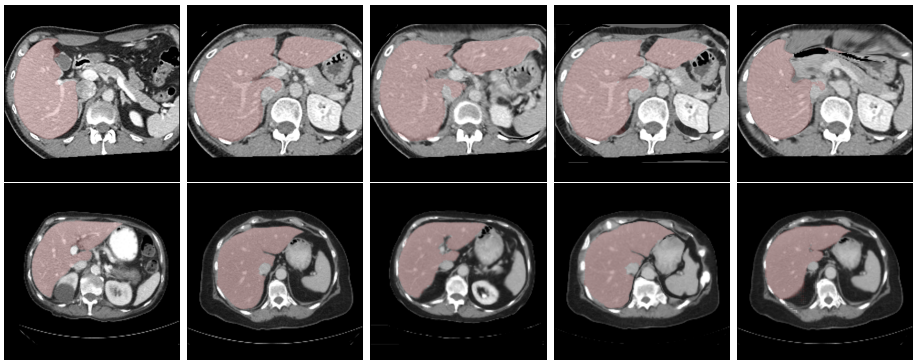

Fig. 2. Exemplary registration results. The images show corresponding slices of the 3D volumes (top: our dataset, bottom: Learn2Reg dataset) with the liver segmentation as red overlay. From left to right: fixed image, moving image, deeds, VoxelMorph, Ours.

In addition, we compared our method to the classical registration algorithm deeds [1] and the deep learning registration approach VoxelMorph [2]. We adapted VoxelMorph from an atlas-based registration to an image to image registration and trained it with our image data. The mutual information loss is weighted with $\lambda = 2.0$. Due to the limited GPU memory, 20 is chosen as the number of intensity bins and a total number of 100 epochs is trained.

We used the average Dice score and the lowest 30% of the liver segmentations as metric for evaluation. Additionally, the average surface distance, the 95% Hausdorff distance, the standard deviation of the Jacobian determinant and the percentage of negative elements of the Jacobian are used as evaluation metrics.

The results for all methods for both datasets are listed in Table 1. An example slice of both networks for all methods are displayed in Fig. 2. Deeds took on average six times longer than the deep learning approaches, with all deep learning approaches taking less than 10 seconds.

4 Discussion and Conclusion

Considering inter-patient registration of the first dataset, our semantically guided registration network outperforms the remaining approaches with Dice score of 0.91 and an average surface distance of 3.90 which corresponds to an improvement of at least +4% and -0.71 voxels, respectively.

In case of the second dataset, our semantically guided registration network also outperforms the other deep learning-based methods by at least +7% for the Dice score and -2.3 voxels for the average surface distance. However, the classical method deeds is the best method for this dataset and has a higher Dice score (+4%) and a smaller surface distance (-0.84 voxels) than our method.

The different results for the two datasets can be explained trough the small number of training images for the second dataset. In general, learning-based methods are strongly dependent of the amount and quality of the dataset which leads to a good generalization of the model.

Overall, our presented approach can overcome the problems of most deep learning-based methods that have difficulties to register large deformations.

References

1. Heinrich MP, Jenkinson M, Brady M, et al. MRF-based deformable registration and ventilation estimation of lung CT. IEEE Trans Med Imaging. 2013;32(7):1239–1248.
2. Balakrishnan G, Zhao A, Sabuncu MR, et al. Voxelmorph: a learning framework for deformable medical image registration. IEEE Trans Med Imaging. 2019;38(8):1788–1800.
3. Ha IY, Wilms M, Heinrich M. Semantically Guided Large Deformation Estimation with Deep Networks. Sensors. 2020;20(5):1392.
4. Sun D, Yang X, Liu MY, et al. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 8934–8943.