2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)
April 13-16, 2021, Nice, France

2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI) | 978-1-6654-1246-9/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISBI48211.2021.9433767

COOPERATING NETWORKS TO ENFORCE A SIMILARITY CONSTRAINT IN PAIRED BUT UNREGISTERED MULTIMODAL LIVER SEGMENTATION

Vincent Couteaux[⋆†] *Mathilde Trintignac*[‡] *Olivier Nempont* [†] *Guillaume Pizaine* [†]
Anna Sesilia Vlachomitrou [†] *Pierre-Jean Valette* [‡] *Laurent Milot* [‡] *Isabelle Bloch* [⋆,⋆⋆]

[⋆]LTCI, Télécom Paris, Institut Polytechnique de Paris, France
[†]Philips Research Paris, France
[‡]Hospices Civils de Lyon, France
[⋆⋆] Sorbonne Université, CNRS, LIP6, Paris, France

ABSTRACT

We propose a method for segmenting two unregistered images from different modalities of the same patient and study at once, while enforcing a similarity constraint between the two segmentation masks. Our method relies on a segmentation network and a registration network, cooperating to get accurate and consistent segmentation masks across modalities, while forcing the segmentor to use all information available. Experiments on a dataset of T1 and T2-weighted liver MRI show that our method enables to get more similar segmentation masks across modalities than manual annotations, without deteriorating the performance (Dice = 0.95 for T1, 0.92 for T2).

Index Terms— Segmentation, Registration, Similarity, Liver, Multimodal Imaging

1. INTRODUCTION

We address the problem of automatic segmentation of two images of different modalities from the same study. Segmentation of multiple modalities are used, for instance, as a preliminary step for automated assessment of hepatic fat fraction or liver tumor burden.

As some modalities are acquired a few minutes apart, or on different machines, organs may not be aligned on the different images. Therefore, every image needs its own segmentation mask of the same organ.

The literature on automatic multimodal segmentation has primarily focused on aligned images, needing only one segmentation, with datasets such as the intervertebral disc dataset [1], in which MRI images are acquired simultaneously. A review of such methods and datasets can be found in [2].

Concurrently, integration of prior information in automatic segmentation gained interest: structure-driven priors with regularization [3], data-driven priors with adversarial learning [4], or knowledge-driven priors by integrating anatomical constraints [5], to name but a few.

Our work fits into the latter class of methods: we want to integrate into the training the knowledge that, as both images represent the same object, segmentation masks should only differ by a certain deformation. We propose to integrate this prior by adding a *relative similarity* constraint as a regularization to the training process. We say that two segmentation masks are similar *relatively to a class of transformations*, if there exists a transformation in this class that maps one mask into the other. The goal of integrating this prior is three-fold: (i) to get more similar segmentations and consequently get more consistent quantitative measurements across modalities; (ii) to limit the effect of bias in the annotations that are specific to a modality (for instance, T2-weighted MRIs often have a poor z-axis resolution, which can cause artefacts in the annotations); (iii) if the organ of interest is not equally easy to segment in the two images (as is often the case, for instance, if one modality is anatomical and the other functional), to teach the network to fetch the relevant information of the easier modality to segment the harder one.

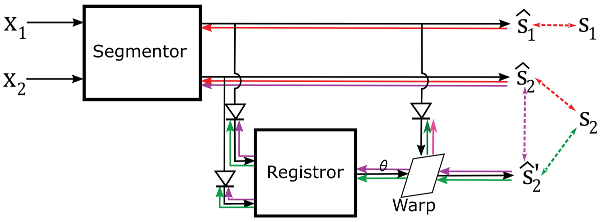

Fig. 1. Overview of our method. Dashed arrows represent loss functions we minimize during training, black arrows represent a forward pass, and colored arrows represent the gradients with respect to the loss of the same color. Diode symbols represent a “stop gradient” operation.

Our method (illustrated in Figure 1, and detailed in Section 2) consists in training simultaneously a *segmentor* network, which takes a pair of images as input and predicts a segmentation for each image, and a *registror* network which takes the predicted masks as input, and returns a transformation that explains the difference between the two masks. Both networks cooperate to minimize the segmentation error and maximize the similarity between both predictions, relatively to the predicted transformation.

This work is motivated by the problem of segmenting the liver in T1 and T2-weighted MRI. Our experiments for this particular problem are detailed in Section 3.2.

2. METHOD

The principle of our method is to create a positive feedback cycle between the segmentor network and the registor network to improve the similarity: as the segmentor outputs more and more similar segmentations, the registror is able to output more precise registrations, which in turn enables to refine the predictions of the segmentor, by learning to fetch the relevant information in the input more accurately.

More precisely, let us consider a dataset of pairs of images (x_1, x_2), and the corresponding annotated segmentation masks (s_1, s_2). We train a segmentor network S such that $(\hat{s_1}, \hat{s_2}) = S(x_1, x_2)$. We restrict ourselves to the setting where a single segmentor is being used, rather than two independent networks, so that it can take into account both images for segmenting each individual image. We simultaneously train a registror R such that $\theta = R(\hat{s_1}, \hat{s_2}) \in \Omega$, where Ω, the set of deformations, controls the acceptable transformation of the object to segment (for instance, if we segment a rigid organ, we can set Ω to be the set of rigid transformations). We denote by $\hat{s_2}' = \theta(\hat{s_1})$ the predicted mask of the first modality, transformed with the deformation predicted by R. The two masks $\hat{s_2}'$ and $\hat{s_2}$ must be as close as possible to s_2, while $\hat{s_1}$ must be close to s_1. As the images of the pair are from different modalities, voxel intensity values do not compare, meaning that simple image-based losses for registration are not suited. In contrast to [6], where unpaired image translation is used, and to [7], where a common representation is learnt, we propose to only register masks to compare the images, which greatly eases the task of registration.

2.1. Loss functions

To create the positive feedback cycle, we define three loss functions:

Segmentation loss: (red path in Figure 1)

$$L_r = L_{bce}(s_1, \hat{s_1}) + L_{bce}(s_2, \hat{s_2})$$

where L_{bce} is the binary cross-entropy loss function.

Registration loss: (green path in Figure 1)

$$L_g = L_{mse}(\hat{s_2}' * f, s_2 * f)$$

where L_{mse} is the Mean Squared Error function, and f is a low-pass filter. We blur the masks before applying L_{mse} in order to soften the edges of the masks, thus avoiding discontinuities, and get more consistent gradients with respect to this loss.

Similarity loss: (pink path in Figure 1)

$$L_p = L_{mse}(\hat{s_2}', \hat{s_2})$$

Both networks are trained to minimize

$$L = \lambda_r L_r + \lambda_g L_g + \lambda_p L_p$$

where $\lambda_r, \lambda_g, \lambda_p$ are hyperparameters. The purpose of L_r and L_g is to train the segmentor and the registror, respectively. Meanwhile, L_p acts as a regularization loss: it constrains the predicted masks $\hat{s_1}$ and $\hat{s_2}$ to differ only by the transformation predicted by the registror, and λ_p is the regularization parameter which controls the trade-off between relative similarity and segmentation precision.

The diode symbols in Figure 1 represent the operation that stops gradients from being back-propagated. Their goal is to prevent the segmentor from being affected by L_g, thus avoiding local minimas where S predicts twice the same mask.

The similarity loss only affects the second channel, and constrains its prediction to be similar to the first channel. This asymmetry between the two modalities is justified if we know that one modality is harder to segment than the other, a situation that often occurs in practice. In this case the first channel receives the easy modality.

2.2. Training and testing

The training is done in three steps: we pre-train the segmentor with the segmentation loss only, the registror with the registror loss and annotation masks as input, and then the whole model with all three losses.

To assess the effect of our method on the similarity of the predicted masks, we define the relative similarity metric of masks s_1 and s_2 relatively to the transformation set Ω:

$$D_S(s_1, s_2, \Omega) = \max_{\tau \in \Omega} Dice(\tau(s_1), s_2)$$

where $Dice(x, y)$ denotes the Dice index of two binary masks x and y. When Ω is the set of smooth and dense deformations, we compute an approximation of D_S by blurring s_1 and s_2 to convexify the problem, and finding the optimal τ by iterative gradient descent:

$$\tau_0 = I_d$$

$$\tau_{k+1} = \tau_k - \alpha \frac{\partial}{\partial \tau_k} ||\tau_k(s_1 * f) - (s_2 * f)||^2$$

where f is a low-pass filter, and α is the gradient descent step. In practice we found that a simple separable $[1, 2, 1]$ was sufficient, for both L_g and this approximation of D_S

3. EXPERIMENTS

3.1. Toy dataset

In order to test the influence of λ_p on the similarity with a controlled setting, we design an experiment with a toy dataset: We generate 240×240 images, where the foreground to segment is made of random blobs, slightly different and randomly translated (Ω is the set of translations of 32 pixels maximum) in s_1 and s_2. To simulate two different modalities, we fill the images with sine patterns where the foreground/background is encoded by the angle in one modality, and frequency in the other (see Figure 2).

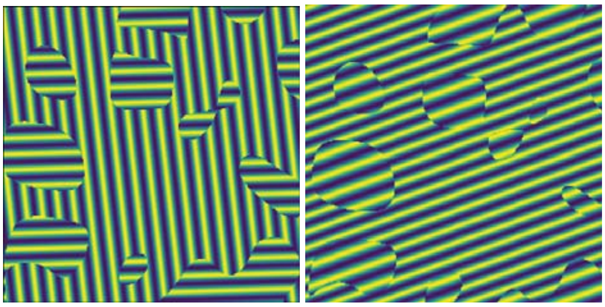

Fig. 2. Pair of images from the toy dataset.

We train 30 models using different values of $\lambda_p \in [0, 1]$, with $\lambda_r = 1 - \lambda_p$ and $\lambda_g = 1$. For each model we compute the similarity metric (relatively to translations) of predictions on a test set, and compare it to the similarity of the ground truth of the same test. Distributions of differences in similarity are shown in Figure 3. We can see a large gain in similarity compared to the ground truth for $\lambda_r = 1 - \lambda_p < 0.5$, which shows that λ_r and λ_p enable to tune the trade-off between the accuracy towards the second modality and the similarity.

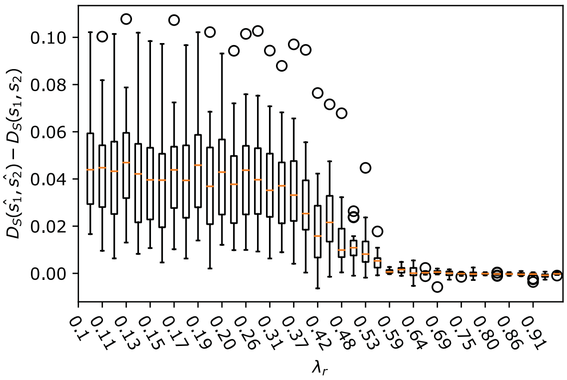

Fig. 3. Effect of λ_r on the toy setting. We compare the similarity (relatively to translations) of the predictions to the similarity of the ground truth. Each box represents the distribution of gain in similarity for the test database, for a particular value of $\lambda_r = 1 - \lambda_p$.

3.2. Liver in T1 and T2-weighted MRI

3.2.1. Data

Our experiments on real data rely on a database containing 88 pairs of T1-weighted and T2-weighted MRIs centered on the liver, coming from 51 patients. The T1 images are acquired at a portal enhancement time. Images have a plutot voxel size of 3mm on the z axis, and 1.5mm for the x and y axes. Every patient's liver has at least one lesion.

Reference segmentation masks are obtained through manual annotations by a radiologist, using in-house 3D interactive tools. Note that due to the low contrast of the liver in T2 images (see Figure 4), as well as the lower resolution along the z axis, the manual annotation of the liver in T2 images is difficult and less accurate than in T1 images.

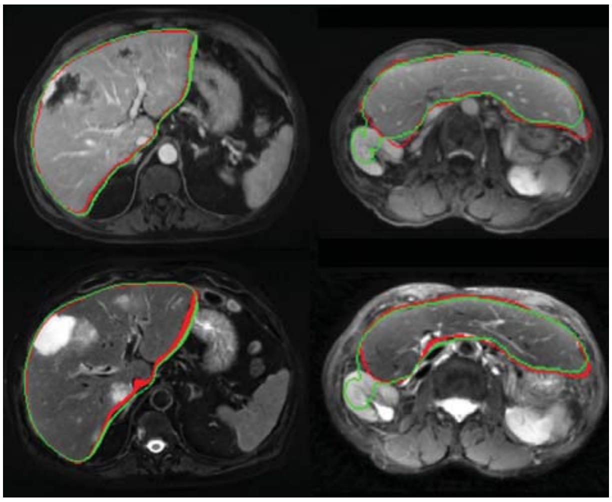

Fig. 4. Two examples from the test dataset. Top: T1-weighted MRI, bottom: T2-weighted MRI. In red, the manual annotation, in green the prediction of our method. The rightmost column shows a particularly challenging case.

3.2.2. Training

We use a 3D U-net architecture with weights provided in [8] for the segmentor, and refer to [8] for the architecture details. As the T1 and T2 images are acquired a few minutes apart, the liver, being made of soft tissue, undergoes a non-rigid deformation, mainly due to breathing. To enforce smoothness, we chose Ω to be the set of elastic deformations defined by displacement vectors on a low-resolution grid, and trilinearly interpolated between vectors.

To estimate this displacement vector field, the registor is a fully-convolutional network, down-sampling the input by a factor 16 (4 blocks of two $3 \times 3 \times 3$ convolutional layers with ReLU activations followed by one $2 \times 2 \times 2$ Max-Pooling layer, each block having 16, 32, 64 and 128 feature maps, respectively, and a $1 \times 1 \times 1$ convolution at the end). We then resample the resulting deformation field to full resolution before warping. Both the warping and the deformation field resampling are done using trilinear interpolation.

We split the dataset keeping 12 pairs for testing, and train the model during 1500 epochs of 100 steps. For memory reasons, we use batches of size 1, and crop inputs into cuboids of random sizes and ratios. We use $\lambda_r = 0.1, \lambda_g = 1, \lambda_p = 1$ and add random intensity shift as data augmentation.

3.2.3. Results

In order to get an idea of the extent of the liver misalignment, we measure the overlap of annotated masks in each pair and get a Dice score of 0.751. To get the relative similarity of the annotated masks, we compute D_S with Ω as described in Section 3.2.2, and obtain a mean similarity of 0.955 for the annotations. A segmentor trained independently (without a registror) achieves predictions with a similarity of 0.954. Predictions of networks trained with our method have a similarity of 0.965. We perform a Wilcoxon signed-rank test to compare the similarity of pairs of our prediction vs. pairs of annotations and get $p = 0.0029$, which tends to show that the difference is not the effect of statistical noise.

We measure performance by comparing the predictions to annotations, and record a Dice of 0.946 for T1 images and 0.918 for T2 images. As a comparison, performance with the segmentor only ($\lambda_p = \lambda_g = 0$) is 0.942 for T1 and 0.897 for T2, whereas a single-input U-net predicting one mask at a time achieves 0.961 for T1 and 0.938 for T2. We recall that the goal of the method was not to improve the performance (as measured by a comparison with the annotations), but rather the similarity of the pairs of predicted mask. For more details and discussion on performance of different training strategies we refer to our other work [9].

T1 predictions warped with the predicted deformation compare with T2 annotations by a Dice of 0.917, which shows the good performance of the registror.

The left column of Figure 4 shows that our method produces accurate segmentations, even when the liver has big lesions near the edge. The right column of Figure 4 presents a challenging case where the liver is abnormally elongated and pushed to the right hand side of the image. No such case is present in the training dataset, leading the network to mistake a part of the left kidney for the liver. However, the consistency of this error in both images highlights the similarity of the segmentations across modalities, which is a desired behavior.

4. CONCLUSION

We proposed a Deep Learning-based method to address the problem of similarity between predictions in paired but unregistered multimodal segmentation. We first tested this method in a controlled setting with a toy dataset, and showed the effect of the regularization parameter, which tunes the trade-off between relative similarity and accuracy. We then applied the method to a real database of liver T1 and T2 MRIs, and showed that the resulting predictions were more similar than the annotations, and than the predictions of naive approaches, without compromising the performances too much. Future work will test how this method enables to stabilize quantitative measurements.

5. REFERENCES

[1] Jose Dolz, Christian Desrosiers, and Ismail Ben Ayed, "IVD-Net: Intervertebral disc localization and segmentation in MRI with a multi-modal UNet," in *International Workshop and Challenge on Computational Methods and Clinical Applications for Spine Imaging*. Springer, 2018, pp. 130–143.

[2] Tongxue Zhou, Su Ruan, and Stéphane Canu, "A review: Deep learning for medical image segmentation using multi-modality fusion," *Array*, vol. 3-4, pp. 100004, 2019.

[3] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez, "ADVENT: Adversarial entropy minimization for domain adaptation in semantic segmentation," *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2512–2521, 2019.

[4] Laurens Samson, Nanne van Noord, Olaf Booij, Michael Hofmann, Efstratios Gavves, and Mohsen Ghafoorian, "I bet you are wrong: Gambling adversarial networks for structured semantic segmentation," in *IEEE International Conference on Computer Vision Workshops*, 2019, pp. 951–960.

[5] Qi Zeng, Davood Karimi, Emily HT Pang, Shahed Mohammed, Caitlin Schneider, Mohammad Honarvar, and Septimiu E Salcudean, "Liver segmentation in magnetic resonance imaging via mean shape fitting with fully convolutional neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 246–254.

[6] Fengze Liu, Jinzheng Cai, Yuankai Huo, Chi-Tung Cheng, Ashwin Raju, Dakai Jin, Jing Xiao, Alan L. Yuille, Le Lu, Chien-Hung Liao, and Adam P. Harrison, "JSSR: A joint synthesis, segmentation, and registration system for 3D multimodal image alignment of large-scale pathological CT scans," *ArXiv:2005.12209*, 2020.

[7] Agisilaos Chartsias, Giorgos Papanastasiou, Chengjia Wang, Scott Semple, David E. Newby, Rohan Dharmakumar, and Sotirios A. Tsaftaris, "Disentangle, align and fuse for multimodal and zero-shot image segmentation," *ArXiv:1911.04417*, 2019.

[8] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway, and Jianming Liang, "Models genesis: Generic autodidactic models for 3D medical image analysis," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 384–393.

[9] Vincent Couteaux, Mathilde Trintignac, Olivier Nempont, Guillaume Pizaine, Anna Sesilia Vlachomitrou, Pierre-Jean Valette, Laurent Milot, and Isabelle Bloch, "Comparing deep learning strategies for paired but unregistered multimodal segmentation of the liver in T1 and T2-weighted MRI," *ArXiv:2101.06979*, 2020.

6. ACKNOWLEDGMENTS

This work has been partially funded by a grant from *Association Nationale de la Recherche et de la Technologie* (#2018/0199).

The authors have no relevant financial or non-financial interests to disclose.

7. COMPLIANCE WITH ETHICAL STANDARDS

This study was performed in line with the principles of the Declaration of Helsinki. This is conform to standard reference methodology MR-004 of the CNIL (France). Approval was granted by the CNIL (Study number 19-188).