Multi-modal unsupervised domain adaptation for deformable registration based on maximum classifier discrepancy

Christian N. Kruse[1], Lasse Hansen[1], Mattias P. Heinrich[1]

[1]Institute of Medical Informatics, Lübeck University
christian.kruse@uni-luebeck.de

Abstract. The scarce availability of labeled data makes multi-modal domain adaptation an interesting approach in medical image analysis. Deep learning-based registration methods, however, still struggle to outperform their non-trained counterparts. Supervised domain adaptation also requires labeled- or other ground truth data. Hence, unsupervised domain adaptation is a valuable goal, that has so far mainly shown success in classification tasks. We are the first to report unsupervised domain adaptation for discrete displacement registration using classifier discrepancy in medical imaging. We train our model with mono-modal registration supervision. For cross-modal registration no supervision is required, instead we use the discrepancy between two classifiers as training loss. We also present a new projected Earth Mover's distance for measuring classifier discrepancy. By projecting the 2D distributions to 1D histograms, the EMD L1 distance can be computed using their cumulative sums.

1 Introduction

Labeled 3D medical images in particular for multi-modal scans are rare due to the necessity of expert knowledge and the large time consumption for labeling. However, datasets containing either a large number of landmarks [1] or anatomical labels [2] are required for training deep learning models for automatic image analysis. Unsupervised domain adaptation for multi-modal or multi-domain images allows using labeled data of one domain to be used on other domains thereby reducing the need for expensive expert-labeled data. In computer-vision, classifier discrepancy for unsupervised domain adaptation has been successfully used for classification and segmentation tasks [3]. This approach requires a metric for comparing the output of two classifiers [4]. Different approaches exist for this discrepancy measures, for example, the Earth Mover's distance (EMD) [5] for 1D cases and adaptations for 2D histograms [6]. These approaches are, however, computationally expensive and based on sensitive hyper-parameters.

1.1 Contribution

We are the first to employ unsupervised domain adaptation for medical image registration using a discrete displacement setting. We train our model for mono-modal registration with strong supervision from pre-computed displacement fields. In a next step we use the discrepancy between two classifiers as a training loss to first maximise the discrepancy by updating the classifier weights and then minimising the discrepancy by updating the feature extractor weights. We further improve over the sliced Wasserstein metric [3] using a novel 2D histogram projected Earth Mover's distance. An early proof-of-concept abstract of this approach on only synthetic MR T1/T2 patches and without instance optimization was published in [7].

1.2 Related work

For supervised multi-modal image registration some recent methods include using a Twin CNN-architecture to predict similarity of patches using aligned multi-modal training data [8] and U-Net like registration with anatomical segmentations [2,9]. In [9] the latter method is extended using a normalized gradient metric. Discrete displacement labeling in deep learning-based registration was proposed in [10] to capture large deformations. In [11] Cycle-GANs were used for unpaired multi-modal segmentation via knowledge distillation. Recently, promising methods for domain adaptation for image classification and multi-modal segmentation have been published in [12].

2 Materials and methods

In this work, we formulate medical image registration as a discrete labelling task to exploit the strengths of domain adaptation for classification tasks. We apply our approach to 2D CT to MR registration using a 21x21 displacement vector resulting in a 441-class classification task.

The training is supervised with pre-computed CT $\rightarrow$ CT and MR $\rightarrow$ MR displacement fields. This pseudo-ground truth was computed using the pdd-net [10]. For the task of registering CT to MR images no supervision is provided.

Our domain adaptation model is shown in Fig. 1: The two input images (240x260 resolution) for registration (fixed and moving image) are first passed to a feature extractor. The feature extractor produces 120x130 feature maps with 24 channels. We then sample a multidimensional tensor from the feature map of the moving image using an identity grid (interpreted as 1D-column vector per image dimension) to which the relative displacement search offsets (interpreted as 1D row vector per image dimension) are added. This yields a resampled feature tensor of size 24x3900x441, where 24 is the number of feature channels, 3900 the number of (downsampled) pixels (quarter resolution: 60x65) and 441 the size of the displacement vector (spatial region of 21×21). This 24x3900x441 feature tensor is then concatenated with the feature map of the fixed image, which is

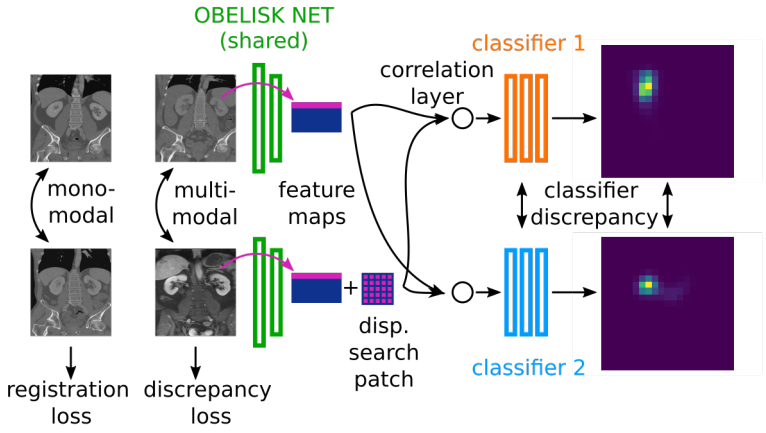

Fig. 1. Domain adaptation approach presented in this work: The model consists of one feature extractor, an OBELISK net, shown in green and two classifiers, shown in red and blue, which yield discrete displacement vectors.

repeated 441 times, and then passed to the classifiers. These classifiers are set up equally except for their random initialization. The classifiers then calculate a smooth metric of both tensors. Each training epoch consists of three steps:

1. train the feature extractor and both classifiers to register CT $\rightarrow$ CT or MR $\rightarrow$ MR (in alternating training epochs) with displacement supervision
2. maximise the discrepancy for cross-modal registration between both classifiers while minimising the classification loss for mono-modal registration by only updating the classifier weights.
3. minimise the classifier discrepancy by updating the feature extractor weights

For feature extraction we employ an OBELISK net [13] with 105k weights. The OBELISK net takes a 240x260 pixel input image (CT or MR) and produces a 120x130 (later down-sampled to 60x65) feature map with 24 channels. For discrete displacement labeling we use two classifiers with 5 blocks of Conv2d, InstanceNorm and PReLU with 33k weights per classifier. The classifiers take two 24x3900x441 feature maps and compute a smooth metric of both inputs. We train the model for 3000 epochs with a learning rate of 0.005 for both the feature net optimizer and the classifier optimizer.

The sliced Wasserstein metric [3] is not invariant to histogram bin/ displacement class permutations and, therefore, not ideally suited for measuring the classifier discrepancy in our approach. In our case we can convert the displacement prediction into 2D spatial probability maps for the x- and y-displacements. Hence, we propose the new projected Earth Mover's distance (p-EMD) as a discrepancy measure for our approach. The EMD for 1D histograms with linear complexity can be exactly solved [5]. Based on the 2D-histogram displacements being close to mono-modal Gaussians we approximate the optimal transport problem by projecting the normalized histograms to 16 lines rotated between

Table 1. Dice scores calculated from 6 labels for different training setups on VISCERAL data. Avg. is the average across all labels. L, S, K and P are the labels for liver, spleen, kidneys and psoas major muscles, respectively, where the two labels each for K and P are averaged.

	CT → CT	MR → MR	CT → MR				
	avg.	avg.	avg.	L	S	K	P
no registration	55.4%	58.4%	50.1%	55.6%	39.6%	48.7%	54.0%
classifier only	77.9%	75.4%	57.4%	76.1%	57.1%	55.7%	50.1%
domain adaptation	76.8%	74.6%	62.3%	77.2%	60.4%	61.5%	56.6%

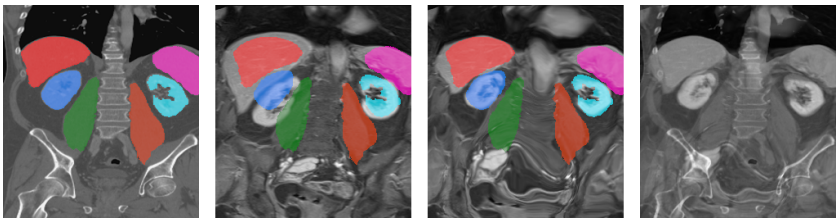

Fig. 2. From left to right: CT image (fixed) with labels for liver (red), spleen (pink), left and right kidney (blue and teal) and left and right psoas major muscle (green and orange), MR image (moving) with CT labels, warped MR image with CT labels and overlay of CT and warped MR image.

0 and 90 degrees with bi-linear interpolation. The L1 distance of the different projections then yields the p-EMD, which matches computationally expensive full EMD calculations almost perfectly and is, at least for our experiments, much more stable than the 2D diffusion distance in [6]. As a post-processing step we perform instance optimization using semi-global matching [14] to improve the alignment.

3 Results and discussion

We test our approach on 2D coronal CT and MR slices from the VISCERAL dataset [15]. We have a total of 9 CT and 9 MR unpaired image slices with six labels for the liver, spleen, kidneys and psoas major muscles. We use all data for training and later also test the cross-modal registration from all 9 CT slices to all 9 MR slices. We employ our training method with supervision on CT and MR in alternating training epochs. We set up the displacement vector to cover displacements of up to 40% of the image width. We test our trained model on mono-modal (CT → CT, MR → MR) and multi-modal (CT → MR) registration and compare the resulting label Dice scores (averaged over all labels) with the initial alignment and the registration with the same model setup but trained without any domain adaptation. The results are shown in table 1. We see that both models improve over initial alignment for mono-modal registration and increase Dice scores from about 55% and 58% to about 77% and 75% for

CT and MR, respectively. The Dice scores for the individual labels for mono-modal registration were also very similar for both training methods and are therefore not shown here. For the cross-modal registration of 81 pairs we see an improved alignment even for the model trained without domain adaption with an increase from 50.1% to 57.4% in Dice score. Our proposed domain adaption approach increases the Dice score by an additional 4.9% points to 62.3%. The label specific Dice scores for multi-modal registration are also shown in table 1. We see that the liver Dice score is only about 1% higher with domain adaptation. The spleen, kidneys and psoas major muscles however benefit stronger from the domain adaptation with a Dice score increase from 3.3% up to 6.5% This shows the benefit of domain adaptation with maximum classifier discrepancy and projected Earth Mover's distance. Future work will focus on an extension to 3D and applying our approach to other multi-modal registration tasks.

Acknowledgement. This work was in part supported by the German ministry of Education and Research (BMBF) within the project Multi-Task Deep Learning for Large-Scale Multimodal Biomed- ical Image Analysis (MDLMA) FKZ 031L0202B.

References

1. Xiao Y, Rivaz H, Chabanas M, et al. Evaluation of MRI to Ultrasound Registration Methods for Brain Shift Correction: The CuRIOUS2018 Challenge. IEEE Trans Med Imaging. 2020 mar;39(3):777–786.
2. Hu Y, Modat M, Gibson E, et al. Weakly-supervised convolutional neural networks for multimodal image registration. Med Image Anal. 2018;49:1–13. Available from: http://www.sciencedirect.com/science/article/pii/S1361841518301051.
3. Lee CY, Batra T, Baig MH, et al. Sliced Wasserstein Discrepancy for Unsupervised Domain Adaptation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2019 mar; p. 10277–10287. Available from: http://arxiv.org/abs/1903.04064.
4. Saito K, Watanabe K, Ushiku Y, et al. Maximum Classifier Discrepancy for Unsupervised Domain Adaptation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2018 dec; p. 3723–3732. Available from: http://arxiv.org/abs/1712.02560.
5. Werman M, Peleg S, Rosenfeld A. A distance metric for multidimensional histograms. Computer Vision, Graphics, and Image Processing. 1985;32(3):328–336. Available from: http://www.sciencedirect.com/science/article/pii/0734189X85900556.
6. Haibin Ling, Okada K. Diffusion Distance for Histogram Comparison. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). vol. 1; 2006. p. 246–253.
7. Heinrich MP, Hansen L. Unsupervised learning of multimodal image registration using domain adaptation with projected Earth Move's discrepancies. Medical Imaging with Deep Learning. 2020 may;Available from: http://arxiv.org/abs/2005.14107.
8. Simonovsky M, Gutiérrez-Becker B, Mateus D, et al. A Deep Metric for Multimodal Registration. In: Ourselin S, Joskowicz L, Sabuncu MR, et al., editors. Medical

Image Computing and Computer-Assisted Intervention - MICCAI 2016. Cham: Springer International Publishing; 2016. p. 10–18.
9. Hering A, Kuckertz S, Heldmann S, et al. Memory-efficient 2.5D convolutional transformer networks for multi-modal deformable registration with weak label supervision applied to whole-heart CT and MRI scans. Int J Comput Assist Radiol Surg. 2019;14(11):1901–1912. Available from: https://doi.org/10.1007/s11548-019-02068-z.
10. Heinrich MP. Closing the Gap between Deep and Conventional Image Registration using Probabilistic Dense Displacement Networks. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2019 jul;11769 LNCS:50–58. Available from: http://arxiv.org/abs/1907.10931.
11. Wolterink JM, Dinkla AM, Savenije MHF, et al. Deep MR to CT Synthesis Using Unpaired Data. In: Tsaftaris SA, Gooya A, Frangi AF, et al., editors. Simulation and Synthesis in Medical Imaging. Cham: Springer International Publishing; 2017. p. 14–23.
12. Dou Q, Liu Q, Heng PA, et al. Unpaired Multi-Modal Segmentation via Knowledge Distillation. IEEE Trans Med Imaging. 2020;39(7):2415–2425.
13. Heinrich MP, Oktay O, Bouteldja N. OBELISK – One Kernel to Solve Nearly Everything: Unified 3D Binary Convolutions for Image Analysis. In: MidlProceedings of the Conference on Medical Imaging with Deep Learning (MIDL). Amsterdam; 2018. p. 9.
14. Hirschmuller H. Accurate and efficient stereo processing by semi-global matching and mutual information. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 2; 2005. p. 807–814 vol. 2.
15. Jimenez-del Toro O, Müller H, Krenn M, et al. Cloud-Based Evaluation of Anatomical Structure Segmentation and Landmark Detection Algorithms: VISCERAL Anatomy Benchmarks. IEEE Trans Med Imaging. 2016 nov;35(11):2459–2475.