Abstract

We propose a semantic similarity metric for image registration. Existing metrics like euclidean distance or normalized cross-correlation focus on aligning intensity values, giving difficulties with low intensity contrast or noise. Our semantic approach learns dataset-specific features that drive the optimization of a learning-based registration model. Comparing to existing unsupervised and supervised methods across multiple image modalities and applications, we achieve consistently high registration accuracy and faster convergence than state of the art, and the learned invariance to noise gives smoother transformations on low-quality images.

1 Introduction

Deformable registration is a fundamental preprocessing tool in medical imaging, where the goal is to find anatomical correspondences between images and derive geometric transformations $\Phi$ to align them. Most algorithmic and deep learning-based methods optimize alignment via a similarity measure $D$ and a $\lambda$-weighted regularizer $R$, combined in a loss function

$$L(I, J, \Phi) = D(I \circ \Phi, J) + \lambda R(\Phi).$$

The similarity metric $D$ assesses the alignment and strongly influences the result. Pixel-based similarity metrics like euclidean distance (MSE) and patch-wise normalized cross-correlation (NCC) are commonly used in both algorithmic \cite{3, 4, 10, 22, 23, 25} and deep learning based \cite{1, 5, 7, 8, 13, 14, 17, 18, 26, 28} image registration. Typically, the similarity measure for a task is selected as the best out of a small set of metrics, with no guarantee that one of the metrics is suitable for the data.

The shortcomings of pixel-based similarity metrics have been studied substantially in the image generation community \cite{29}, where the introduction of deep similarity metrics approximating human visual perception has improved the generation of photo-realistic images \cite{6, 12}. As registration models are generative models \cite{7}, we expect these similarity metrics to improve registration as well. Current attempts at using learned similarity metrics for image registration require ground truth transformations \cite{11} or limit the input to the registration model \cite{17}.

We propose a data-driven similarity metric for image registration based on the alignment of semantic features. We learn semantic filters of our metric on the dataset, use it to train a registration model, and validate our approach on three biomedical datasets of different image modalities and applications. Across all datasets, our method achieves consistently high registration accuracy, outperforming even metrics utilizing supervised information. Our models converge faster and learn to ignore noisy image patches, leading to smoother transformations on low-quality data.
2 A deep similarity metric for image registration

To align areas of similar semantic value, we propose a similarity metric based on the agreement of semantic feature representations of two images. Semantic feature maps are obtained by a feature extractor to be tuned on a surrogate segmentation task. To capture alignment of both localized, concrete features, and global, abstract ones, we calculate the similarity at multiple layers of abstraction.

Concretely, given a set of feature-extracting functions $F^l : \mathbb{R}^{\Omega \times C} \rightarrow \mathbb{R}^{\Omega_l \times C_l}$ for $L$ layers, we define

$$
\text{DeepSim}(I \circ \Phi, J) = \frac{1}{L} \sum_{l=1}^{L} \frac{1}{|\Omega_l|} \sum_{p \in \Omega_l} \frac{\langle F^l_p(I \circ \Phi), F^l_p(J) \rangle}{\|F^l_p(I \circ \Phi)\| \|F^l_p(J)\|},
$$

where $F^l_p(J)$ denotes the $l^{th}$ layer feature extractor applied to image $J$, at spatial coordinate $p$. It is a vector of $C_l$ output channels, and the spatial size of the $l^{th}$ feature map is denoted by $|\Omega_l|$. The metric is influenced by the neighbourhood of a pixel, as $F^l$ composes convolutional filters with increasing receptive area of the composition. Note that the formulation via cosine similarity is similar to the classic NCC metric, which can be interpreted as the squared cosine-similarity between two zero-mean vectors of patch descriptions.

**Feature extraction** To aid registration, the functions $F^l(\cdot)$ should extract features of semantic relevance for the registration task, while ignoring noise and artifacts. We achieve this by training the feature extractor on a supplementary segmentation task, as segmentation models excel at learning relevant kernels for the data while attaining invariance towards non-predictive features like noise. The obtained convolutional filters act as feature extractors for DeepSim, see also Figure 1.

3 Experiments

We compare registration models trained with DeepSim to the baselines MSE, NCC, NCC$_{\text{sup}}$ (NCC with supervised information [5]), and VGG (a VGG-net based metric common in image generation and similar to our method [12]). Figure 1 shows our model architecture. For both registration and segmentation we use U-nets [21]. The registration network predicts the transformation $\Phi$ based on two images $I, J$. A spatial transformer module [15] applies $\Phi$ to obtain the morphed image $I \circ \Phi$. The loss is given by Eq. 1 we choose the diffusion regularizer for $R$ and tune hyperparameter $\lambda$ on the validation sets.

To show that our approach is applicable to a large variety of registration tasks, we validate it on three 2D and 3D datasets of different image modalities: T1-weighted Brain-MRI scans [9, 16], human blood cells of the Platelet-EM dataset [20], and cell-tracking of the PhC-U373 dataset [15, 21]. Each dataset is split into a train, validation, and test section.
### Registration accuracy & convergence

We measure the mean Sørensen Dice coefficient on the unseen test-set in Figure 2 and test for statistical significance of the result with the Wilcoxon signed rank test for paired samples. Our null hypothesis for each similarity metric is that the model trained with DeepSim performs better. We test for a statistical significance levels of $p^* = 0.05$, $p^{**} = 0.01$, $p^{***} = 0.001$. We further measure the effect size with Cohen’s $d$, and label the metrics accordingly in Figure 2. Models trained with our proposed DeepSim rank as the best on the Brain-MRI and Platelet-EM datasets, with strong statistical significance. On the PhC-U373 dataset, all models achieve high dice-overlaps of $> 0.97$. DeepSim converges faster than the baselines, especially in the first epochs of training.

### Qualitative examples & transformation grids

We plot the fixed and moving images $I, J$ and the morphed image $I \circ \Phi$ for each similarity metric model in Figure 3a, and a more detailed view of a noisy patch of the Platelet-EM dataset in Figure 3b. The transformation is visualized by grid-lines, which have been transformed from a uniformly spaced grid. On models trained with the baselines, we find strongly distorted transformation fields in noisy areas of the images. In particular, models trained with NCC and NCC$^{\text{sup}}$ produce very irregular transformations, despite careful tuning of the regularization-hyper-parameter. The model trained by DeepSim is more invariant towards the noise.

## 4 Discussion & Conclusion

Registration models trained with DeepSim achieve high registration accuracy across multiple datasets, leading to improved downstream analysis and diagnosis in medical applications. The consistency of our proposed metric makes testing multiple traditional metrics unnecessary; instead of empirically determining whether MSE or NCC captures the characteristics of a data-set best, we can use DeepSim to learn the relevant features from the data.

The analysis of noisy patches in Figure 3b highlights a learned invariance to noise. The pixel-based similarity metrics are distracted by artifacts, leading to overly-detailed transformation fields. DeepSim does not show this problem. While smoother transformation fields can be obtained for all metrics by strengthening the regularizer, this would negatively impact the registration accuracy of anatomically significant regions. Accurate registration of noisy, low-quality images allows for shorter acquisition time and reduced radiation dose in medical applications.

DeepSim is a general metric, applicable to image registration tasks of all modalities and anatomies. Beyond the presented datasets, the good results on low-quality data let us hope that DeepSim will improve registration accuracy in the domains of lung CT and ultrasound, where details are hard to identify, and image quality is often poor. We further emphasize that the application of DeepSim is not limited to deep learning. Algorithmic image registration uses a similar optimization framework, where a similarity-based loss is minimized via gradient descent-based methods. DeepSim can be applied to drive algorithmic methods, improving their performance by aligning deep, semantic feature embeddings.
Acknowledgments and Disclosure of Funding

The authors acknowledges support by the Lundbeck Foundation grant number R218-2016-883.

Broader impact

The broader impact of our work is defined by the numerous applications of medical image registration. Common applications are in neuroscience [5], CT-imaging of lungs and abdomen [26], as well as for fusion and combination of multiple modalities [11].

The deep learning approach to image registration utilized in this work can achieve impressive results across a wide variety of tasks, but this often comes at the cost of training models on specialized hardware for extensive periods. This energy-intensive workload may raise carbon emissions, the primary contributor to climate change [2]. We hope that by presenting a method for learning a semantic similarity metric from the data, we make excessive testing of other loss functions unnecessary. This can reduce the amount of model configurations to be tested in the development of deep learning methods, contributing to a lower environmental impact of the image registration community.

References


