DeepSim: A Semantic Approach to Image Registration Evaluation

Abstract

This paper introduces a novel semantic similarity metric designed for image registration. Current metrics, such as Euclidean distance or normalized cross-correlation, primarily focus on aligning intensity values, which presents challenges when dealing with low contrast or noise. Our approach utilizes learned, dataset-specific features to guide the optimization of learning-based registration models. In comparisons with existing unsupervised and supervised methods across various image modalities and applications, our method demonstrates consistently superior registration accuracy and faster convergence. Additionally, its learned noise invariance results in smoother transformations on lower-quality images.

1 Introduction

This paper delves into the significant area of deformable registration, an essential preprocessing step in medical imaging. The primary objective is to ascertain anatomical correspondences between images and determine geometric transformations, denoted as Φ, for their alignment. The majority of algorithmic and deep learning-based techniques achieve alignment by optimizing a similarity measure, D, and a λ-weighted regularizer, R, which are combined to form a loss function:

$$L(I, J, \Phi) = D(I \circ \Phi, J) + \lambda R(\Phi). \tag{1}$$

The alignment is critically evaluated by the similarity metric, D, which significantly impacts the final outcome. Common pixel-based metrics, such as Euclidean distance (MSE) and patch-wise normalized cross-correlation (NCC), are used in both algorithmic and deep learning approaches to image registration. Typically, a similarity measure for a particular task is selected from a small set of metrics, with no certainty that any of them is suitable for the data.

The limitations of pixel-based similarity metrics have been extensively studied in the image generation field, where the adoption of deep similarity metrics, designed to emulate human visual perception, has enhanced the generation of highly realistic images. Because registration models are also generative, we anticipate that employing these similarity metrics could also improve registration results. However, current methods that use learned similarity metrics for image registration require ground truth transformations, or they restrict the input to the registration model.

We propose a data-driven similarity metric for image registration that relies on aligning semantic features. Our metric uses learned semantic filters specific to the dataset, which are then used to train a registration model. We have validated our method using three biomedical datasets characterized by varying image modalities and applications. Across all datasets, our approach achieves consistently high registration accuracy, even outperforming metrics that use supervised information. Our models also demonstrate quicker convergence and learn to overlook noisy image patches, leading to more consistent transformations on lower-quality data.

.

2 A Deep Similarity Metric for Image Registration

To align areas with comparable semantic content, we propose a similarity metric based on the consensus of semantic feature representations between two images. These semantic feature maps are generated by a feature extractor, trained through a surrogate segmentation task. To capture the alignment of both localized, specific features and more abstract, global ones, we compute the similarity across multiple layers of abstraction.

Given a set of feature-extracting functions, $F_l : \mathbb{R}^{\Omega \times C} \to \mathbb{R}^{\Omega_l \times C_l}$, for L layers, we define:

$$DeepSim(I \circ \Phi, J) = \sum_{l=1}^{L} \frac{1}{|\Omega_l|} \sum_{p \in \Omega_l} \frac{F_l(I \circ \Phi)_p \cdot F_l(J)_p}{\|F_l(I \circ \Phi)_p\| \|F_l(J)_p\|} \tag{2}$$

where $F_l(J)_p$ denotes the l-th layer feature extractor applied to image J at spatial coordinate p. It is represented as a vector of C_l output channels, and the spatial size of the l-th feature map is denoted as $|\Omega_l|$. The metric is influenced by the pixel's neighborhood, since F_l uses convolutional filters with an expanding receptive area. Note that the formulation, using cosine similarity, mirrors the classic NCC metric, which can be interpreted as the squared cosine-similarity between two zero-mean patch description vectors.

To improve registration, the functions $F_l(\cdot)$ should extract features that are semantically relevant to the registration task, while ignoring noise and artifacts. This is achieved by training the feature extractor on an additional segmentation task, since segmentation models excel at learning pertinent kernels while also achieving invariance to features like noise that are not predictive. The convolutional filters obtained act as feature extractors for DeepSim.

3 Experiments

We evaluated registration models trained with DeepSim against baseline metrics such as MSE, NCC, NCCsup (NCC using supervised information), and VGG (a VGG-based metric used in image generation, similar to our approach). The model architecture is shown in Figure 1. For both registration and segmentation, we used U-nets. The registration network predicts the transformation Φ based on two input images, I and J. The spatial transformer module applies Φ to obtain the morphed image $I \circ \Phi$. The loss function is as in Eq. 1; we chose the diffusion regularizer for R and fine-tuned the hyperparameter λ on the validation sets.

To demonstrate the broad applicability of our method across various registration tasks, we assessed it using three datasets of both 2D and 3D images with different image modalities: T1-weighted Brain-MRI scans, human blood cells from the Platelet-EM dataset, and cell tracking from the PhC-U373 dataset. Each dataset was divided into training, validation, and testing subsets.

4 Results

Table 1: Quantitative comparison of similarity metrics. Stars indicate p-test significance level. Effect size given by Cohen's d.

	Brain-MRI	Platelet-EM	PhC-U373
MSE	0.70	0.98‡	0.98
NCC	0.71‡	0.98‡	0.98
NCCsup	0.72‡	0.98‡	0.98
VGG	0.71‡	0.98‡	0.98
DeepSim	0.75	0.99	0.99

‡ indicates p<0.001 statistical significance with effect size > 0.8.

Registration Accuracy Convergence: We evaluated the mean Sørensen-Dice coefficient on the unseen test set (Table 1) and tested the statistical significance of the results using the Wilcoxon signed-rank test for paired samples. The null hypothesis for each similarity metric was that the model trained with DeepSim would perform better. Statistical significance levels were set at $p^* = 0.05$, $p^{**} = 0.01$, and $p^{***} = 0.001$. Additionally, we used Cohen's d to measure the effect size. Models trained with our proposed DeepSim were ranked highest on both the Brain-MRI and Platelet-EM datasets, exhibiting strong statistical significance. In the PhC-U373 dataset, all models achieved a high dice-overlap exceeding 0.97. DeepSim converged faster than the baseline models, particularly during the initial training epochs.

Qualitative Examples Transformation Grids: We display the fixed and moving images, I and J, along with the transformed image $I \circ \Phi$, for each similarity metric model in Figure 2(a), and a more detailed view of a noisy patch from the Platelet-EM dataset in Figure 2(b). The transformation is shown using grid-lines, which were transformed from an evenly spaced grid. We observed considerably distorted transformation fields in noisy image areas in models trained with the baselines. Specifically, models trained with NCC and NCCsup demonstrated highly irregular transformations, despite the careful adjustment of the regularization hyperparameter. The model trained with DeepSim showed greater invariance to noise.

5 Discussion and Conclusion

Registration models trained with DeepSim show substantial registration accuracy across multiple datasets, which improves downstream medical analysis and diagnostics. The reliability of our proposed metric reduces the need for testing multiple traditional metrics. Instead of experimentally determining whether MSE or NCC best captures the properties of a dataset, DeepSim can be used to learn the appropriate features from the data.

The analysis of noisy patches in Figure 2(b) highlights an inherent resistance to noise. Pixel-based similarity metrics are influenced by artifacts, leading to excessively detailed transformation fields, which DeepSim does not exhibit. Although smoother transformation fields can be achieved for all metrics by increasing the regularizer, this would negatively affect the registration precision of anatomically important areas. Accurate registration of noisy, low-quality images allows for shorter acquisition times and reduced radiation in medical applications.

DeepSim is a general metric that can be applied to image registration across all modalities and anatomies. Beyond the presented datasets, good results on low-quality data suggest that DeepSim could improve registration accuracy in lung CT and ultrasound imaging, where details are difficult to identify, and image quality is often compromised. Furthermore, DeepSim is not restricted to deep learning; algorithmic image registration follows a comparable optimization structure where similarity-based loss is minimized through gradient descent methods. Applying DeepSim in algorithmic methods can improve their performance by aligning deep, semantic feature embeddings.

6 Broader Impact

The widespread applications of medical image registration significantly amplify the broader impact of our work. Some of the typical applications include neuroscience, CT imaging of the lungs and abdomen, as well as the fusion and combination of different modalities.

The use of deep learning for image registration, while capable of achieving remarkable outcomes across many different applications, often necessitates the training of models using specialized hardware over extended periods. This energy-intensive task may raise carbon emissions, which are a major contributor to climate change. By introducing a method that learns a semantic similarity metric directly from data, we hope to eliminate the need for excessive testing of other loss functions. This can reduce the number of model configurations tested during the development of deep learning methods, thus contributing to a lower environmental impact within the image registration community.