Discriminative Segmentation-based Evaluation through Shape Dissimilarity

Ender Konukoglu, Ben Glocker, DongHye Ye, Antonio Criminisi, and Kilian M. Pohl

***Abstract*—Segmentation-based scores play an important role in the evaluation of computational tools in medical image analysis. These scores evaluate the quality of various tasks, such as image registration and segmentation, by measuring the similarity between two binary label maps. Commonly these measurements blend two aspects of the similarity: pose misalignments and shape discrepancies. Not being able to distinguish between these two aspects, these scores often yield similar results to a widely varying range of different segmentation pairs. Consequently, the comparisons and analysis achieved by interpreting these scores become questionable. In this paper we address this problem by exploring a new segmentation-based score, called normalized Weighted Spectral Distance (nWSD), that measures only shape discrepancies using the spectrum of the Laplace operator. Through experiments on synthetic and real data we demonstrate that nWSD provides additional information for evaluating differences between segmentations, which is not captured by other commonly used scores. Our results demonstrate that when jointly used with other scores, such as Dice's similarity coefficient, the additional information provided by nWSD allows richer, more discriminative evaluations. We show for the task of registration that through this addition we can distinguish different types of registration errors. This allows us to identify the source of errors and discriminate registration results which so far had to be treated as being of similar quality in previous evaluation studies.**

***Index Terms*—Evaluation, Accuracy Assessment, Image Registration, Image Segmentation, Shape Dissimilarity, Overlap Measures, Spectral Distance, Shape Dissimilarity, Laplace Operators.**

I. Introduction

EVALUATION of computational tools in medical image analysis is an important task. Widespread application of these tools in different research fields, their deployment on commercial systems, their use in advanced analysis tasks and the amount of basic research focusing on developing new tools emphasize the need of sound evaluation methodologies. This need not only arises for understanding which algorithm performs better on a specific dataset. It is also crucial for devising unit tests for commercial systems, understanding algorithm limitations for clinical use, detecting failures in applications involving large amount of data and interpreting analysis results correctly.

Though very important, evaluations for most analysis tools are not straightforward. The main difficulty is the lack of ground truth or gold standard. One particular tool is very striking in this regard: *image registration* [1]–[3]. Registration is defined as determining the coordinate transformation between two images that aligns the corresponding anatomical points. It is used for a wide range of purposes such as fusing images of different modalities of the same anatomy [4], studying spatiotemporal dynamics [5] and performing large cohort studies [6].

Evaluating a registration method is defined as assessing the accuracy of the coordinate transformation computed by the method. In theory, this assessment can be done simply by comparing the computed transformation with the real transformation between the images. However, this is precisely the point where it becomes difficult. The "real" transformation between two arbitrary images is usually unknown, and thus, ground truth for evaluation is inaccessible.

Despite the difficulty in its assessment, many analyses rely on registration. Their outcome and correctness heavily depend on the accuracy of the computed coordinate transformation. This issue has been discussed, for example, in the context of voxel-based [7] and deformation-based morphometry [8]. In 2003, Crum *et al.* in [9] remarked: "Clinical studies whose results rely heavily on registration techniques of questionable validity should be treated with suspicion."

In order to circumvent the lack of ground truth, scientists resort to using indirect or sparse methods for accuracy assessment. Different approaches include using synthetically generated transformations [10], [11], using sparse set of landmarks to quantify alignment errors [11]–[13] and quantifying mathematical properties of the computed coordinate transformation [11], [13]–[16]. Although used in various studies, these approaches are either too application-specific, in the case of landmarks and synthetic transformations, or not indicative of the registration accuracy. The last group of approaches, which is also the most widely used one, uses segmentations of corresponding structures [17], [18].

Segmentation-based approaches for assessing accuracy of a given coordinate transformation is based on the fact that the correct transformation between two images would align the corresponding anatomical structures perfectly. These methods thus quantify the quality of the coordinate transformation by measuring the discrepancies between the corresponding segmentations after registration. Although segmentation-based methods do not directly quantify the registration accuracy (in terms of mm displacements of corresponding anatomical points), they are the most generally applicable and the most popular group of evaluation strategies. This popularity is mainly due to: i) creating manual segmentations of structures is often easier and less sensitive to noise than annotating

E. Konukoglu (corresponding author, email: ender.konukoglu@gmail.com), B. Glocker and A. Criminisi are with Microsoft Research Ltd., Cambridge, UK.

D.H. Ye and K.M. Pohl are with the Department of Radiology, University of Pennsylvania, USA.

landmarks, ii) existence of publicly available imaging studies that include scans and associated expert segmentations, iii) segmentations in some sense provide a “dense sampling of landmarks along the boundary” (given that the exact correspondence of such landmarks between reference and floating image is unknown), which enable the computation of a wide array of measurements, such as the overlap agreement between regions, and iv) one of the major applications of registration is segmentation via label propagation and therefore, measuring the registration accuracy via segmentations is closely aligned to the target application.

Although very popular and useful, segmentation-based scores that are commonly used in the literature have limitations, [17], [19]. The one that is tackled in this article is that scoring functions, such as Dice’s similarity coefficient (DSC) [20] and surface distance, are often not discriminative enough when it comes to certain differences between segmentations. They measure the differences by blending two sources of imperfections: i) pose misalignments (linear) and ii) shape discrepancies (nonlinear). However, they cannot discriminate these two sources and as a result, for a large class of visually very different segmentation pairs, these functions return very similar scores. Such an example is shown in Figure 1. In the context of registration, this means that coordinate transformations of different qualities might not be correctly discriminated, which undermines the assessment. Here, we focus on this issue and address it by exploring a scoring function that ignores pose misalignment and only measures shape discrepancies.

Specifically, we present a score of *shape dissimilarity*, called *normalized Weighted Spectral Distance (nWSD)*. nWSD is a normalized score (in the interval $[0, 1]$) that quantifies the amount of discrepancy between two shapes by using their Laplace spectra. In doing so, it enriches segmentation-based evaluation by providing an additional measurement that cannot be solely captured using other scores. Here, we define nWSD and analyse its properties. Through different experiments with synthetic and real data, we demonstrate that nWSD i) can capture and quantify shape differences independently from pose misalignments, and ii) can complement existing scores leading to more discriminative and richer evaluation.

We first demonstrate the limitations of commonly used segmentation-based scores. We do so by constructing in Section II a very simple database of segmentations where popular scores, such as DSC, are not able to discriminate between visually very different segmentation pairs. In Section III, we provide some technical background on Laplace operators and their role in shape analysis. We then present nWSD, which can provide the necessary discrimination. For the sake of brevity and focus, we omit the theoretical analysis of our score, which is described in [21]. Instead we discuss a series of synthetic examples to underline the properties and the advantages of nWSD for the purposes of this article. Based on these examples, we then propose a two dimensional evaluation system, which jointly uses nWSD along with an overlap score, namely DSC. In Section IV we apply the two dimensional system on real data for assessing the quality of 306 registrations cross aligning MRI brain scans of 18 different subjects. The experiment highlights the additional information provided by nWSD and the use of this richer assessment in the registration scenario. In particular the results demonstrate that nWSD allows us to interpret differences in DSC, where higher not always means better.

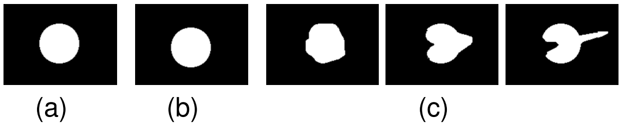


Fig. 1: Example images from the synthetic database. Image in (a) is the reference disc of radius 15 mm, followed by four perturbed versions of this reference in (b) and (c). The first image shown in (b) is a simple translation of (a) by 3 mm. The remaining three shown in (c) are nonlinear deformations of the reference with varying magnitude and amount of nonlinearity. By construction, the DICE scores between the reference and all the perturbed ones are identical.

II. Studying Common Scoring Functions

Commonly used segmentation-based scoring functions quantify the differences between two label maps taking into account: i) misalignments due to incorrect pose and ii) shape (geometry) discrepancies. The scores are applied in the evaluation of registration methods for *indirectly* measuring the quality of anatomical correspondences between the aligned images. This evaluation is indirect as it rather measures errors of overlap and resemblance of corresponding regions, than the errors in actual point correspondences. Popular parameter-free measures are DSC, symmetric mean surface distance (SMSD), symmetric root-mean-square error over surface distance (SRMS), Hausdorff distance (HD), volume similarity (VS) [17], and other statistics based on true/false positives and negatives such as overlap score (OS). Their popularity partly lies in their ease of implementation and intuitive meaning. While HD and SRMS are more sensitive to shape differences by responding to the largest errors, DSC, SMSD, OS, and VS are more robust to outliers and segmentation errors.

All these segmentation-based scores have limitations when it comes to distinguishing shape differences, whether subtle or substantial. The robust ones, such as DICE and VS, cannot discriminate misalignments due to incorrect pose from mismatches in shape even in the case when the shapes are significantly different. The more sensitive measures such as HD and SRMS, essentially measure the dissimilarity between the boundaries of the segmentations in terms of locations but not in terms of their overall geometry. As a result, when applied to evaluate registration algorithms, these measures may yield similar scores to substantially different registration outcomes. We demonstrate these shortcomings on a synthetic database of 2D label maps.

Our synthetic database consists of a reference label map showing a disc of radius 15 mm in an image of 200×200 pixels with a resolution of 0.5 mm, see Fig. 1(a). By randomly perturbing the reference, we created 250 other segmentations. One can imagine these new segmentations to represent different possible registration results with respect to the reference.

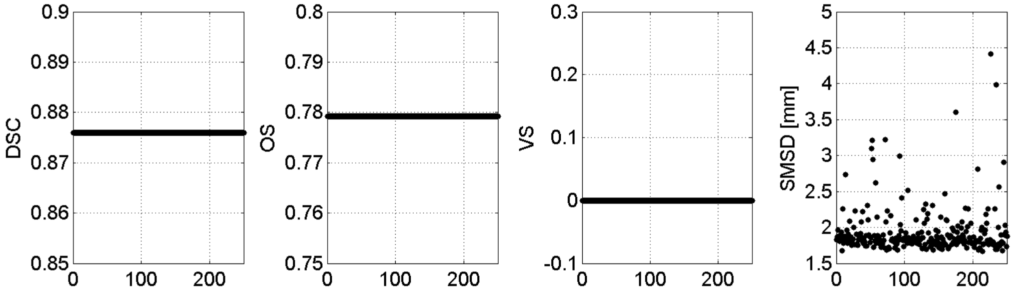
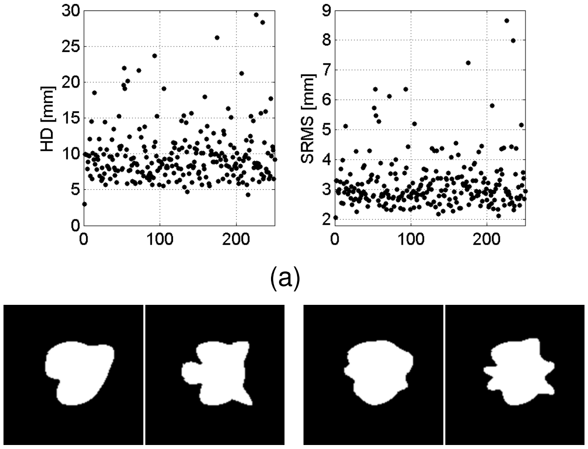


Fig. 2: Different overlap similarity scores applied to the synthetic database (examples shown in Figure 1). The graphs show different similarity scores between the reference label map and each of the 250 perturbed images. The x-axis in each graph is the index of the perturbed image. Notice, the commonly used segmentation-based accuracy scores are unable to properly capture the substantial shape variation in the constructed dataset.



(a)

(b) HD[mm]: 8.5

(c) SRMS[mm]: 2.4

Fig. 3: The HD and SRMS scores for the synthetic dataset. (a) The graphs plot the HD and SRMS scores between the reference image and the perturbed images. We notice that HD and SRMS show higher variations throughout the synthetic database compared to the scores in Figure 2. However, this dispersion does not correspond to shape differences. (b) Two perturbed segmentations whose HD distances to the reference disc are very similar, both 8.5 mm. (c) Two other perturbed segmentations whose SRMS scores are very similar, both 2.4 mm. HD and SRMS scores are unable to acknowledge the shape differences.

The first perturbed segmentation is a simple translation of the reference by 3 mm, therefore has exactly the same shape but a misalignment with respect to pose, see Fig. 1(b). The remaining 249 segmentations are created by deforming the reference shape using transformations with varying magnitude and amount of nonlinearity. As a result, they all have different shapes than the reference, as the samples shown in Fig. 1(c). As an additional constraint, the dataset is constructed such that the DICE scores between each perturbed image and the reference shape is identical. We note the wide variations of the sample shapes shown in Figure 1. Now, we analyse the commonly used scores using this dataset.

We first compute DSC, SMSD, OS and VS between the reference image and all the other perturbed images. Figures 2 plots these scores for each perturbed segmentation. As expected DSC, OS and VS are exactly the same for all the images. The same, although not shown here, is actually also true for other measures, such as various statistics based on true/false positives. They do not capture the shape differences and as a result they cannot distinguish between errors in pose and shape discrepancy. The SMSD score shows some variation between different segmentations however: i) this variation is very small, i.e. 200 of 250 images are within interval $[1.6, 2.0]$ mm and ii) there is no discrimination with respect to shape.

We also obtain measurements on the synthetic database using HD and SRMS scores, although these scores are normally not used due to their high susceptibility to outliers. The HD and SRMS scores are shown in Figure 3(a). As expected, the dispersions for these scores are much higher throughout the dataset compared to the previously shown scores. However, the dispersions do not necessarily correlate with the shape differences between the segmentations. Figure 3(b) illustrates this issue with an example. The two segmentations shown in this figure have very similar, identical up to the first floating point, HD scores with respect to the reference disc. The HD score fails to identify the substantial shape differences between the segmentations. We observe a similar behaviour for SRMS with respect to the segmentations of Figure 3(c). These examples show that, in addition to their high susceptibility to outliers, HD and SRMS are unable to capture certain shape discrepancies. Furthermore, we would also like to point out that just as we constructed a dataset which has identical DSC scores with respect to the reference shape, one can also construct a dataset that would have identical HD or SRMS scores with respect to the reference.

These simple tests demonstrate that DSC, SMSD, SMRS, OS, and VS are generally not discriminative enough to allow a distinction between misalignments due to pose and shape differences. Even in the case where the shape differences are substantial these scores will not be able to identify them. Considering that many registration evaluation studies [16]–[18] are based on these measures, and different algorithms are ranked by considering a few percent differences in their scores, the limitation of the segmentation-based scores is critical.

It is crucial that the measures discussed in this section are considered in combination with others. Examples are overlap distance [22] or PCA [14]. However, these latter measures are either also not discriminative or require training data for which statistics of the residual transformation are not trivial to obtain.

III. Spectral Shape Dissimilarity

In this section, we explore the use of the spectrum of Laplace operators to define a shape dissimilarity score *normalized Weighted Spectral Distance – nWSD*. We show that nWSD captures and quantifies shape differences and offers a solution to the limitations of existing scores.

We begin this section by first briefly providing the necessary background on Laplace operators, their spectra and their role in shape analysis, Section III-A. Then we present in Section III-B the nWSD score and its properties that make it useful for measuring shape differences. In Section III-C we experimentally analyse nWSD and demonstrate its advantages for the problem segmentation-based evaluations. We further show that jointly using nWSD with DSC yields more discriminative power than either of the scores alone. Finally, we briefly discuss implementation details of nWSD and the choices of its parameters in Section III-D.

A. Spectrum of Laplace Operator

Laplace operators and their spectra have been studied in mathematics and theoretical physics for a long time [23]. Their introduction in computational shape analysis is however, rather recent [24]. In this first part, we give a brief overview of Laplace operators to equip the reader with the necessary background. We specifically focus on their role in shape analysis. For a more thorough discussion of these topics we refer the reader to [23], [25], [26] and [24].

We denote an object (an anatomical structure) as a closed bounded domain $\Omega \subset \mathbb{R}^d$ with piecewise smooth boundaries in the d-dimensional Euclidean space. With respect to images or volumes, Ω corresponds to the region outlined by the segmentation (or the label map). Now let $\mathbb{F}_\Omega \triangleq \{f | f : \Omega \rightarrow \mathbb{R}\}$ be the space of real-valued functions on Ω and $\mathbb{D}_\Omega$ the space of twice differentiable functions in $\mathbb{F}_\Omega$, then the Laplace operator $\Delta_\Omega : \mathbb{D}_\Omega \rightarrow \mathbb{F}_\Omega$ for $f \in \mathbb{D}_\Omega$ with respect to Ω is defined as

$$\Delta_\Omega f \triangleq \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2} f,$$

where $x \triangleq \{x_1, \dots x_d\}$ represent the spatial coordinates of $\mathbb{R}^d$. The importance of the Laplace operator for shape analysis arises from the fact that the eigenvalues and the eigenfunctions of this operator contain information on the intrinsic geometry of the object [23], [27], [28]. An intuitive analogy (in 2D) is to consider a drum membrane that has the same shape as the object. Then, the eigenvalues of the Laplace operator defined on the object correspond to the fundamental frequencies of vibration of the membrane during percussion. These frequencies depend on the shape of the drum and as such the eigenvalues depend on the shape of the object. Mathematically, the eigenvalues and the eigenfunctions of Δ_Ω are the solutions of the Helmholtz equation with Dirichlet type boundary conditions[1], [23],

$$\Delta_\Omega f + \lambda f = 0, \ \forall \mathbf{x} \in \Omega \text{ and } \ f(\mathbf{x}) = 0, \ \forall \mathbf{x} \in \partial\Omega,$$

where $\partial\Omega$ denotes the boundary of the object and $\lambda \in \mathbb{R}$ is a scalar. The eigenvalue-eigenfunction pairs $\{(\lambda_n, f_n)\}_{n=1}^{\infty}$ that satisfy this equation form an infinite set. Furthermore, the ordered set of eigenvalues is a positive diverging sequence $0 < \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n \leq \dots$. This infinite sequence is called the Dirichlet spectrum of Δ_Ω, which we refer simply as the "spectrum". In addition, each component of the spectrum is called a "mode", e.g. λ_n is the n^{th} mode of the spectrum.

As we mentioned above, the spectrum contains information on the intrinsic geometry of objects. Mathematically, this is given by the *heat-trace*, which in $\mathbb{R}^d$ is

$$Z(t) \triangleq \sum_{n=1}^{\infty} e^{-\lambda_n t} = \sum_{s=0}^{\infty} a_{s/2} t^{-d/2+s/2}, \quad t > 0, \quad (1)$$

where t is formally related to a time variable in a heat diffusion system [29]. The coefficients of the polynomial expansion, $a_{s/2}$, are the components carrying the geometric information. These coefficients are given as sums of volume and boundary integrals of some local invariants of the shape, [26], [27], [30], [31]. For instance, as given in [30], the first three coefficients are:

$$a_0 = \frac{1}{(4\pi)^{d/2}} V_\Omega,$$
$$a_{1/2} = -\frac{1}{4(4\pi)^{d/2-1/2}} S_\Omega,$$
$$a_1 = -\frac{1}{6(4\pi)^{d/2}} \int_{\partial\Omega} \kappa d\partial\Omega,$$

where V_Ω is the volume, S_Ω is the surface area (circumference in 2D) and κ is the mean (geodesic) curvature on the boundary of Ω. The functional relationship between the eigenvalue sequence and the coefficients $a_{s/2}$ as given by the Equation (1) relates the spectrum to the intrinsic geometry. This "spectrum-geometry" link makes the Laplace spectrum important for the computational study of shapes.

In addition to the spectrum-geometry link, the spectrum of the Laplace operator has two other properties that make it useful for shape analysis, [23]: i) eigenvalues are invariant to isometric transformations and ii) eigenvalues change continuously with the deformations applied to the boundary of the object. The first property shows that the eigenvalues capture the fact that isometric transformations do not alter the shape but only the location and the orientation of an object. The second property, on the other hand, states that there is a continuous link between the differences in eigenvalues and the difference in shape. This continuous link is a key property that makes the spectrum useful in quantifying shape differences.

Unfortunately, it has also been shown that there exist non-congruent shapes that have exactly the same spectrum, called *isospectral* shapes [32]. Therefore, theoretically the spectrum does not uniquely identify shapes. However, as stated in [24],

[1] Here we are only interested in the Dirichlet type. Please refer to [23] for other types.

practically this does not cause a problem mostly because the constructed isospectral shapes in 2D and 3D are rather extreme examples with nonsmooth and nonconvex boundaries. Furthermore, for dimensions less than four, it is not even clear whether there exist continuous deformations that do not modify the spectrum while changing the shape [33].

Although the spectrum-geometry link has been known for a long time, this link has not been explored for computational analysis of shapes until recently. Inspired by the properties of the spectrum, Reuter *et al.* in [24] proposed a shape descriptor, called *shape-DNA*, based on the eigenvalue sequence. For a given shape Ω, it is defined as the vector composed of the first (smallest) N modes of the spectrum of the associated Laplace operator (i.e. the operator defined on Ω) : $[\lambda_1, \lambda_2, \ldots, \lambda_N]$. Using shape-DNA, authors in [34] and [35] analysed anatomical structures and showed the potential of the spectrum as a descriptive feature vector. They were able to capture the shape differences between distinct objects and use shape-DNA for the purposes of classification, recognition and statistical analysis.

In the context of segmentation-based evaluation, the common measures discussed in Section II mainly use the spatial information extracted from the segmentations. For instance, DSC computes the spatial overlap between the corresponding segmentations. These measures thus combine pose and shape differences in one score. Now, the Laplace spectra present other opportunities. As a representation, using exactly the same input as the other measures, the spectrum extracts information on the intrinsic geometry from the segmentation. Therefore, a scoring function that can quantify the difference between spectra of two objects can also be used as a measure of shape dissimilarity. As a result, such a scoring function alleviates the limitations of existing scores.

B. Normalized Weighted Spectral Distance - nWSD

In order to make use of the shape information contained in the Laplace spectra, we need to define a score or a distance that quantifies the difference between the spectra of two objects. Defining such a *shape-distance* however, is a challenging task due to the diverging nature of the eigenvalue sequence.

A first approach is presented in [24], where a distance is defined as the Euclidean distance between shape-DNAs of objects. Although this distance can be useful for certain cases, it has some important drawbacks [36]. The Euclidean distance between shape-DNAs: i) is extremely sensitive to the descriptor size N while the choice of this parameter is arbitrary, ii) cannot be defined over the entire spectrum, iii) is dominated by the differences at higher modes of the spectrum even though these modes are not necessarily more informative about the intrinsic geometry and iv) cannot be normalized and therefore, it is not trivial to use in conjunction with other scores that have different ranges, such as DSC which is in the interval $[0, 1]$. These problems limit the use of the Euclidean distance in practice.

Here we present an alternative definition, which overcomes the difficulties posed by the diverging nature of the spectrum. In order to keep the presentation focused on the problem of measuring discrepancies between segmentations we only provide the definitions and briefly describe the properties. The derivations and the detailed theoretical analysis of the following definitions, in a more general framework, are presented in [21].

To define our shape dissimilarity score, we first create a theoretically sound spectral distance that can be normalized to the $[0, 1]$ interval. The weighted spectral distance (WSD) for two closed bounded domains with piecewise smooth boundaries, $\Omega_\lambda, \Omega_\xi \subset \mathbb{R}^d$, whose spectra are given as the sequences $\{\lambda_n\}_{n=1}^{\infty}$ and $\{\xi_n\}_{n=1}^{\infty}$ respectively, is defined as

$$\rho(\Omega_\lambda, \Omega_\xi) \triangleq \left[\sum_{n=1}^{\infty} \left| \frac{1}{\lambda_n} - \frac{1}{\xi_n} \right|^p \right]^{\frac{1}{p}}, \tag{2}$$

where p is a positive scalar such that $p > d/2$. Unlike the Euclidean distance between shape-DNAs, WSD is defined over the entire eigenvalue sequence and the difference at each mode uses $1/\lambda_n$ and $1/\xi_n$ rather than λ_n and ξ_n. The basic theoretical properties of WSD are:

(i) for $p > d/2$, WSD exists for any two closed bounded domains with piecewise smooth boundaries, i.e. the infinite sum in the definition is guaranteed to converge to a finite value
(ii) WSD satisfies the triangular inequality making it a pseudometric and
(iii) WSD has a multi-scale aspect with respect to p in the sense that increasing p lowers the sensitivity of WSD with respect to shape differences at finer scales, i.e. with respect to geometric differences at local level such as thin protrusions or small bumps.

Based on the first property of WSD, we can now define the normalized score of shape dissimilarity, which we call *normalized weighted spectral distance (nWSD)*, as

$$\overline{\rho}(\Omega_\lambda, \Omega_\xi) \triangleq \frac{\rho(\Omega_\lambda, \Omega_\xi)}{\mathbf{W}(\Omega_\lambda, \Omega_\xi)} \in [0, 1) \tag{3}$$

where $\rho(\Omega_\lambda, \Omega_\xi)$ is mapped to the $[0, 1)$ interval using the shape-dependent normalization factor

$$\mathbf{W}(\Omega_\lambda, \Omega_\xi) \triangleq \left\{ C + K \cdot \left[\zeta\left(\frac{2p}{d}\right) - 1 - \left(\frac{1}{2}\right)^{\frac{2p}{d}} \right] \right\}^{\frac{1}{p}}.$$

$\zeta(\cdot)$ represents the Riemann zeta function [37], and C and K are the shape based coefficients given as

$$\begin{aligned} C &\triangleq \sum_{i=1,2} \left[\frac{d+2}{d \cdot 4\pi^2} \cdot \left(\frac{B_d \hat{V}}{i}\right)^{\frac{2}{d}} - \frac{1}{\mu} \cdot \left(\frac{d}{d+4}\right)^{i-1} \right]^p \\ K &\triangleq \left[\frac{d+2}{d \cdot 4\pi^2} \cdot \left(B_d \hat{V}\right)^{\frac{2}{d}} - \frac{1}{\mu} \cdot \frac{d}{d+2.64} \right]^p, \\ \hat{V} &\triangleq \max\left(V_{\Omega_\lambda}, V_{\Omega_\xi}\right) \text{ and } \mu \triangleq \max\left(\lambda_1, \xi_1\right), \end{aligned}$$

where B_d is the volume of the unit ball in $\mathbb{R}^d$. nWSD inherits the properties of WSD except being a pseudometric. Furthermore, being confined to $[0, 1)$, nWSD also allows us to i) easily use the shape dissimilarity in combination with scores quantifying other types of differences between objects, such as DSC, and ii) compare dissimilarities of different pairs of shapes which is of practical importance in the setting of registration evaluation.

An important theoretical property of the nWSD score is that it is defined over the entire eigenvalue sequence. In practice, however, we can only compute a finite number of eigenvalues and therefore, can only approximate nWSD. Considering this, we also define the finite approximations of nWSD using the smallest N eigenvalues as

$$\overline{\rho}^N(\Omega_\lambda,\Omega_\xi) \triangleq \frac{\rho^N(\Omega_\lambda,\Omega_\xi)}{\mathbf{W}(\Omega_\lambda,\Omega_\xi)} \in [0,1), \tag{4}$$

which has diminishing asymptotic errors $\lim_{N\to\infty}|\overline{\rho}(\Omega_\lambda,\Omega_\xi) - \overline{\rho}^N(\Omega_\lambda,\Omega_\xi)| = 0$. Furthermore, $\overline{\rho}^N(\cdot,\cdot)$ can accurately approximate nWSD only using a few number of modes, which makes nWSD useful in practice.

The invariance properties of the eigenvalues is the other very important property of nWSD. Since the eigenvalues do not change with respect to isometric transformations, e.g. rotation and translation, the $\overline{\rho}(\cdot,\cdot)$ does not change with respect to isometric transformations applied to the objects. As a result of these invariance properties the nWSD score focuses solely on the shape differences between objects becoming truly complementary to other scores discussed in Section II.

The nWSD score allows us to use the shape information encoded via the Laplace spectra for measuring shape discrepancies between binary label maps.

C. Experimental Analysis of nWSD using Synthetic Images

We now explore nWSD experimentally and analyse its properties from the viewpoint of segmentation-based evaluation by reviewing a series of experiments based on synthetic data. Specifically, we confirm in Section III-C1 the ability of our measure to capture shape differences that are missed by the scores studied in Section II. Furthermore, we perform experiments that demonstrate nWSD's invariance to isometric transformations (Section III-C2) and its continuous relationship with respect to deformations (Section III-C3). These findings serve as a motivation in III-C4 to combine nWSD with DSC resulting in a rich quantification of differences between two binary label maps. Consequently, in the scenario of registration, this yields a more discriminative assessment of registration quality than possible by either score alone.

1) Discriminating Shape Differences: We start our experiments with the dataset of Section II. Following the same procedure as before, we compute nWSD scores between the reference, i.e. a disc of 15 mm radius, and each of the 250 perturbed segmentations, where the first one is a translation of the reference by 3 mm (see also Figure 1). In Figure 4(a) we plot these scores for each perturbed segmentation along with some example images that lie at different bands of the nWSD score. We make the following important observations:

- Considering the value ranges we see that the dispersion of the nWSD score for this dataset is substantially larger than for other scores used in Section II. This shows that nWSD provides a much higher level of discrimination for the segmentations considered in this dataset.

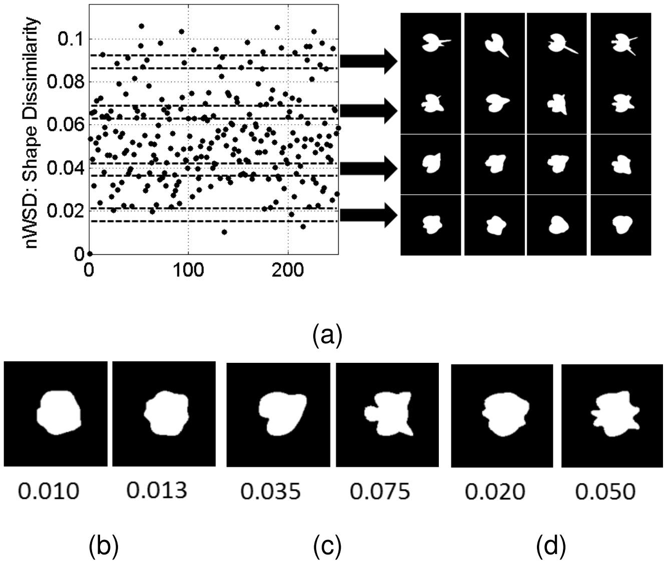


Fig. 4: nWSD scores for the synthetic dataset described in Section II. (a) The graph shows the nWSD scores between the reference segmentation and each of the 250 perturbed segmentations, where x-axis is the image index. The 16 images on the right are some examples of perturbed segmentations corresponding to different bands of nWSD score (same row = similar scores). Note that segmentations with similar scores are visually more similar than ones with very different scores. (b) The two perturbed segmentations with the second and the third lowest nWSD scores with respect to the reference disc. Although the shape differences are subtle they are captured by nWSD. (c) The same images as in Figure 3(b). The difference between their nWSD scores is relatively large considering the maximum and minimum values of nWSD seen in plot (a). nWSD captures the difference between these shapes, while HD does not. (d) The same images as in Figure 3(c), where we see a similar behaviour: the shape difference that is not differentiated by SRMS is captured by nWSD.

- The first image in the dataset, which is simply a translation of the reference segmentation, received the lowest nWSD score, $nWSD = 7.5 \times 10^{-14}$.
- Observing the example segmentations and the corresponding bands of the nWSD score shown in Figure 4(a), we notice that the ordering of shapes with respect to nWSD is visually meaningful. Segmentations that receive similar nWSD scores with respect to the reference have indeed visually comparable discrepancies. It is remarkable that all these segmentations yield *identical* DSC scores with respect to the reference, as shown in Fig. 2.
- The images with the second and third lowest nWSD score are shown in Figure 4(b), from left to right respectively. We note that these segmentations have fairly subtle shape differences compared to the reference. Yet nWSD is able to capture these differences.
- Figure 4(c) and (d) show the pairs of segmentations

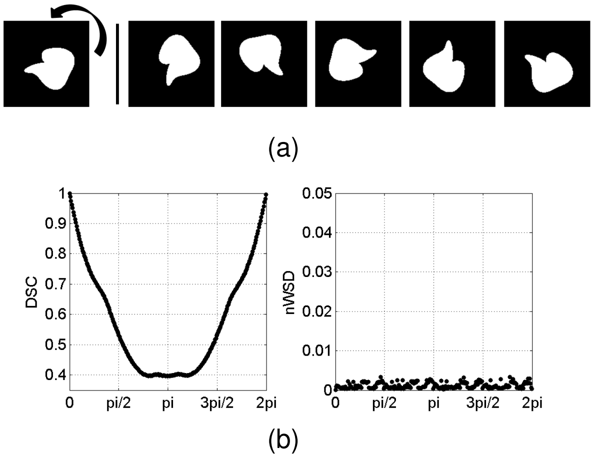


Fig. 5: Invariance of nWSD to isometric transformations. (a) A synthetic dataset of 250 perturbed images obtained by rotating a reference segmentation with angles varying in $[0, 2\pi]$. The left most image shows the reference while the remaining are examples from the perturbed ones. (b) The graph on the left plots the DSC scores between the reference and the perturbed images with respect to the angle of rotation. The graph on the right similarly plots the nWSD score.

that were earlier used in Figures 3(b) and (c), where we have illustrated the limitations of HD and SRMS. Below each segmentation we also give their nWSD scores with respect to the reference. We see that nWSD discriminates between these segmentations while not having the drawbacks of HD and SRMS.

In summary, while other segmentation-based scores fail to capture shape differences in this dataset, nWSD captures the differences and provides a visually meaningful discrimination between different segmentations. We also notice that nWSD does not capture all the differences between two segmentations, i.e. misalignments due to incorrect pose. This is due to its invariance to isometric transformations, which we will explore in the next section. Before proceeding to this analysis, we would like to point out that this invariance is precisely why nWSD is able to provide additional information to the other scores and enriches segmentation-based evaluation.

2) The Source of Extra Information: Invariance to Isometric Transformations: As we have mentioned in Section II, scoring functions, such as DSC, measure the differences between two label maps by blending discrepancies arising from misalignments due to pose and actual shape mismatches. Due to this, they are unable to distinguish between simple translations and substantial shape differences. nWSD only focuses on the shape differences providing that extra information. It achieves this as a consequence of its invariance to isometric transformations. In this section, we demonstrate this invariance of nWSD through a simple example.

For simplicity, we only focus on rotations however similar results can be produced with translations. We constructed a synthetic dataset which consists of a reference segmentation, shown in Figure 5(a) left most image, and rotations of this segmentation with angles varying in $[0, 2\pi]$. Examples of the rotated reference segmentation are shown in Figure 5(a). We then compute DSC and nWSD scores between the reference and the rotated segmentations. Figure 5(b) shows these scores with respect to the angle of rotation. We observe that, as expected, DSC changes with respect to the angle of rotation successfully capturing the misalignment due to pose. The nWSD score varies slightly within the small interval $[0, 0.003]$, meaning that the shape similarities between the reference and the rotated segmentations are almost perfect. In theory, the score should exactly be 0 for all the rotated segmentations. The small deviation from 0 is due to image discretization artifacts, which slightly change the shape.

This experiment demonstrates that the score values obtained via nWSD purely quantify the shape differences, in other words nonlinear discrepancies between two segmentations. As such, nWSD can point out the shape differences without being affected by misalignments due to incorrect pose. This provides a richer understanding of the discrepancies between segmentations and a better interpretation of other scores.

3) Continuity with Respect to Deformations: Another important property of the spectrum mentioned in Section III is that the eigenvalues change continuously with respect to deformations applied to an object's boundary [23]. Here, we use the notion of continuity in the mathematical sense [37]. The continuous relation between the deformations and the spectrum is also inherited by the nWSD score, i.e. the score depends continuously on the deformations. We demonstrate this with a simple example. We start from a reference segmentation – a disc of radius 15 mm – and protrude the boundary of this reference in a continuous manner to create 160 perturbed segmentations. Some examples of these perturbed segmentations are shown in Figure 6. We then computed the nWSD score between the reference and each perturbed segmentation. The graph shown in Figure 6 plots the nWSD scores versus the maximum extent of the protrusion.

Figure 6 shows that nWSD depends continuously on the extent of the protrusion. This continuous relation is especially interesting in segmentation-based scoring for the problem of assessing registration quality because it relates the amount of deformation to the magnitude of the spectral distance.

4) 2D Accuracy Measure: Combining nWSD with DSC Score: The properties that we demonstrated above make nWSD a useful and complementary segmentation-based evaluation score. Motivated by its properties, in this section we propose a two dimensional scoring system where one dimension is DSC, quantifying the overall differences between the segmentations through spatial overlap, and the other dimension is nWSD, focusing on the shape discrepancies. We note that instead of DSC we could have also used one of the other scores studied in Section II.

We follow the same procedure as in the previous sections and make use of a synthetic dataset generated from a reference segmentation. The reference segmentation for this experiment is chosen as the slightly more complicated structure shown in Figure 7(a) left most image. Starting from this reference, we have generated 500 perturbed segmentations using transformations with varying degrees of nonlinearity and magnitude. In

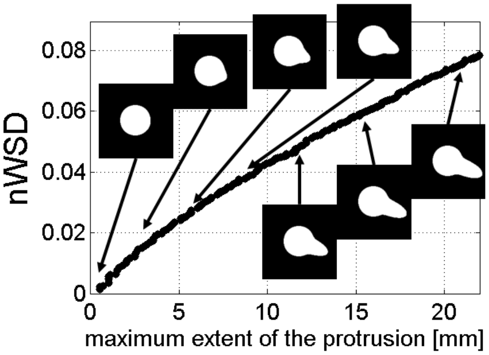


Fig. 6: Change of nWSD with respect to continuously growing deformation. The synthetic dataset in this experiment consists of a reference disc and 160 perturbed images. The perturbed images are constructed by protruding the boundary of the reference in a continuous manner and taking snapshots at different points. The graph plots the nWSD score between the reference and the perturbed images. Some of the perturbed images are shown on the graph pointing to their respective nWSD score.

addition, for the last 100 perturbed segmentations, we have only used isometric transformations, i.e. rotations and translations. Some examples of the perturbed segmentations are shown in Figure 7(a). We then computed the DSC and nWSD scores between the reference and each perturbed segmentation. In Figure 7(b) we plot the nWSD scores. As expected there is a large dispersion among the first 400 images. Furthermore, the last 100 images receive very low scores.

In Figure 7(c) we combine DSC with nWSD in a two dimensional evaluation score. Each point in this graph corresponds to a different perturbed segmentation, the dots representing the ones constructed using nonlinear transformations and the crosses representing the ones constructed using isometric transformations. We observe that there are many segmentations that have very similar DSC scores but different nWSD scores. First of all, this 2D dispersion allow us to further discriminate between these segmentations, which would not have been possible by using only DSC. Furthermore, we can now interpret the sources of the discrepancies as measured by DSC: whether the discrepancy is due to pose misalignments or shape mismatches. Lastly, using this system we can better compare segmentation pairs that yield slightly different DSC values and interpret the difference correctly. By looking at nWSD values for these segmentations, we can conclude whether higher DSC score corresponds to a truly better alignment of the segmentations or if the increase in DSC is coming at the expense of altering the shape. We will elaborate on this idea further in Section IV in the context of an intersubject registration scenario.

In summary, we see that the 2D score (DSC,nWSD) provides richer information on the discrepancies between segmentations. A "good" registration in this plot would lie on the bottom-right corner close to the point (DSC, nWSD)= $(1, 0)$. At this point we can ensure that the structures are not only overlaid well but also that their shapes are similar. Furthermore, comparing segmentations to the template we can conclude that the segmentations with higher DSC score is the result of truly a better alignment if it also has similar or lower nWSD score. If it has a higher nWSD score then this points out that the increase in DSC came at the expense of increasing shape differences.

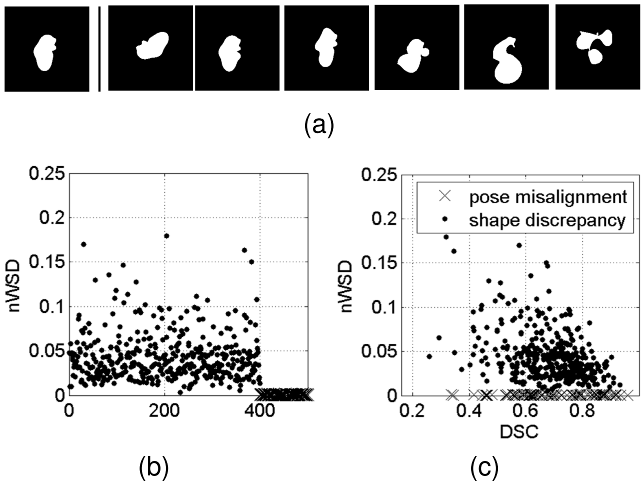


Fig. 7: Joint use of nWSD with DSC. (a) Examples from the synthetic dataset constructed for this experiment. The left most image is the reference image from which 500 other segmentations are constructed by perturbing the reference via deformations of varying magnitude and amount of nonlinearity. The last 100 perturbed images are the result of applying isometric transformations to the reference. (b) The graph shows the nWSD score between the reference and the perturbed images. (c) The graph shows DSC vs. nWSD scores for the perturbed images with respect to the reference. Each point in the graph represents a perturbed image. The crosses correspond to the images which are isometric transformations and the dots correspond to the images which are nonlinear deformations of the reference. This 2D accuracy system provides richer information regarding discrepancies between the perturbed and the reference segmentations. For the same DSC score now we can identify the source of the discrepancy, i.e. whether pose or shape. A good registration in this plot lies on the bottom-right corner where we can not only ensure that the corresponding structures in the aligned images are well overlaid, but also guarantee that their shapes are similar.

D. Implementation Details

There are two different aspects of the numerical computation of the nWSD score. The first one concerns the computation of the Laplace spectra for each segmentation. Most existing numerical methods [24], [38] for computing eigenvalues of the Laplace operator in a volume or on a surface can be used to compute nWSD. For the experiments provided in this article, we choose to use the basic finite difference scheme using the natural image grid (see for instance Chapter 2 of [38] for further details). Our main argument in choosing this method is to avoid additional steps, such as volumetric

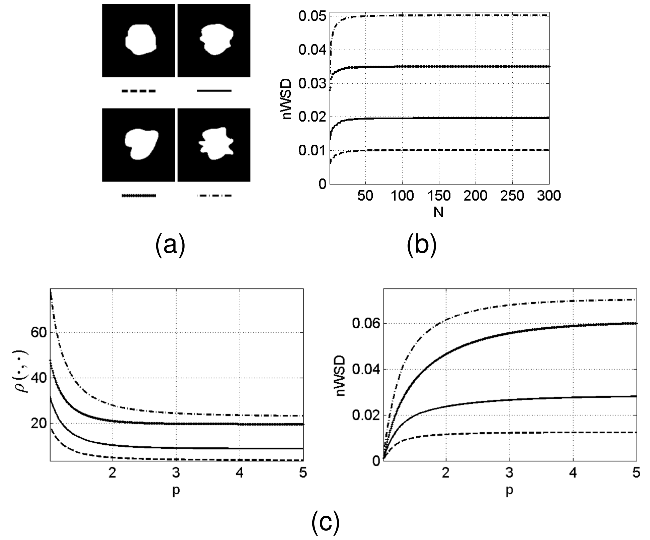


Fig. 8: Evolution of nWSD with respect to the parameters N and p. The line types shown below the images of (a) are used for plotting the corresponding results in (b) - (c). The nWSD scores are computed between each of segmentations shown in (a) and the reference segmentation of Section II shown in Figure 1(a). We see that the convergence for all the segmentations is rapid with respect to the number of modes N. Furthermore, the choice of p do not alter the ordering for these segmentations.

mesh construction, but use exactly the same inputs as other segmentation-based scores. For a segmentation Ω represented as binary image in a rectangular grid we compute the discrete Laplacian operator using the central finite-difference approximation of the Δ_Ω operator. This step yields a sparse matrix which we then solve using Arnoldi's method [39] as implemented in MATLAB®.

The second numerical aspect in computing nWSD is the choice of the parameters p, the norm type, and N, the number of modes. In order to provide the reader an intuition we plot in Figures 8 the change of nWSD (and $\rho(\cdot,\cdot)$) with respect to these parameters. We choose four segmentations from the synthetic dataset used in Section II and compute their nWSD scores with respect to the reference segmentation (the disc of 15 mm radius shown in Figure 1(a)) using different p and N. In Figure 8(a) we show these images where the small strips below the images displays the line style each image corresponds to in the accompanying plots. In Figure 8(b) we plot the evolution of nWSD scores with respect to N (setting $p = 1.5$). We notice that the nWSD scores converge rapidly as N increases and do not change after $N = 50$. Although not shown here, the same convergence holds for any pair of segmentation and in 3D as well. Therefore, the choice of N is not a very crucial parameter for the computation of $nWSD$ as long as it is a fairly large number, such as $N > 50$. In all the experiments shown in this article, whether 2D or 3D, we have used $N = 200$.

In Figure 8(c) we plot the evolution of $\rho(\cdot,\cdot)$ (see Eqn. 2) and nWSD with respect to different p, for the same four segmentations (setting $N = 200$). As we can see in the plot to the left, $\rho(\cdot,\cdot)$ increases with decreasing p. This is as expected since, as mentioned in Section III-B, as p decreases $\rho(\cdot,\cdot)$ becomes sensitive to finer scale shape discrepancies showing higher distances. On the other hand, in the formulation of nWSD given in Eqn. 3 we notice that p also affects the normalization factor $\mathbf{W}$. Integrating this effect as well, we see in the plot to the right that nWSD increases with increasing p and then converges. The important point to notice in these plots is that the order of the curves do not change with p. the exact value of p is application dependent and should be chosen keeping in mind two points. First, low values of p will emphasize the very fine scale differences, which might be due to noise. Therefore, for having a robust score one should not choose p too low. Second, too high values of p might loose details which can be important to distinguish between segmentations. In our experiments we empirically found that the values $p = 1.5$ in 2D and $p = 2.0$ in 3D provide good discrimination while being robust to noise.

In summary, we described a new score that exploits the properties and advantages of spectral representations to measure discrepancies between segmentations. We showed on synthetic images that nWSD is a sound and useful scoring function. It allows us to further distinguish pairs of segmentations with similar DSC score and better interpret the score differences as it ignores pose changes and only measures shape differences.

IV. Inter-Subject MRI Brain Registration

We now perform a series of real data nonlinear registration experiments to underline the benefits of nWSD in practice. For these experiments we use the publicly available dataset IBSR (Internet Brain Segmentation Repository [2]). This dataset includes MR brain scans (T1) of 18 healthy subjects along with manual delineations of 43 structures – subcortical and cortical – for each scan. In [17] the IBSR dataset has been used for a comparative study of different registration algorithms. The comparisons were based on various segmentation-based evaluation scores using the manual delineations. Regarding the behaviour of different scores, the authors state in their results: "Target, union and mean overlap measures for volumes and surfaces (and the inverse of their false positive and false negative values) all produced results that are almost identical if corrected for baseline discrepancies." In other words, different measures did not provide extra information for discriminating different algorithms. Here, we demonstrate that nWSD indeed provides additional information to commonly used segmentation-based scores on the same dataset. We show that when used jointly with DSC, nWSD yields a much richer discrimination between different registrations.

In our experiments we use a single registration algorithm and compare the outcome of different registrations, i.e. different source and target images. For this purpose we employ the diffeomorphic demons algorithm [40] implemented within the ITK library (http://www.itk.org). We cross registered each

[2] http://www.cma.mhg.harvard.edu/ibsr

image in the IBSR dataset to the remaining 17, adding up to 306 non-rigid registrations in total. Each registration is run for 50 iterations and 3 resolution levels. The images have been skull-stripped prior to registration, and the non-rigid registration is initialized with an affine alignment.

After registering the scans, for each pair of source-target images we align the manual segmentations of the corresponding structures using the transformations computed by the registration algorithm. For all cases, we compute the DSC, SMSD, and nWSD scores between the aligned segmentations of four selected structures: right ventricle, caudate, thalamus, and putamen. In Figure 9, we plot these scores for each registration in the 2D coordinate systems (DSC,SMSD) and (DSC,nWSD). In each graph every point corresponds to a different registration problem, i.e. different source-target image pair.

The plots demonstrate the large variation of each score across the dataset and the relationships between different scores. Observing these plots we note the following:

- Figures 9(a)-(c) show that DSC and SMSD are highly correlated for the corresponding structures, i.e. Pearson's correlation coefficients are $r = -0.95, -0.98, -0.98$ respectively. This means that these two scores do not provide different information regarding the quality of the registration with respect to these structures, which is inline with the results given in [17].
- The combination of DSC and nWSD for the same structures shows much less correlation ($r = -0.59, -0.35, -0.48$). There is a large number of registrations that have very similar DSC and different nWSD scores, and vice-versa. This shows for this experiment that nWSD provides additional information to DSC and the proposed 2D scoring system has a higher disriminative power than each score alone.
- The plots in (d) show that the variations of all three scores across the dataset are larger for ventricles than for the other structures. We see that the correlation between DSC and SMSD is still high in this case but slightly lower than the previous cases, i.e. $r = -0.85$. Correlation between DSC and nWSD is also higher, $r = -0.82$. This high correlation is largely due to registrations yielding bad values in all scores. In fact, if we only consider registrations that achieve DSC score higher than 0.7 then the correlation between DSC and nWSD drops yielding $r = -0.33$. A similar behaviour to a lesser extent is also apparent in (a).

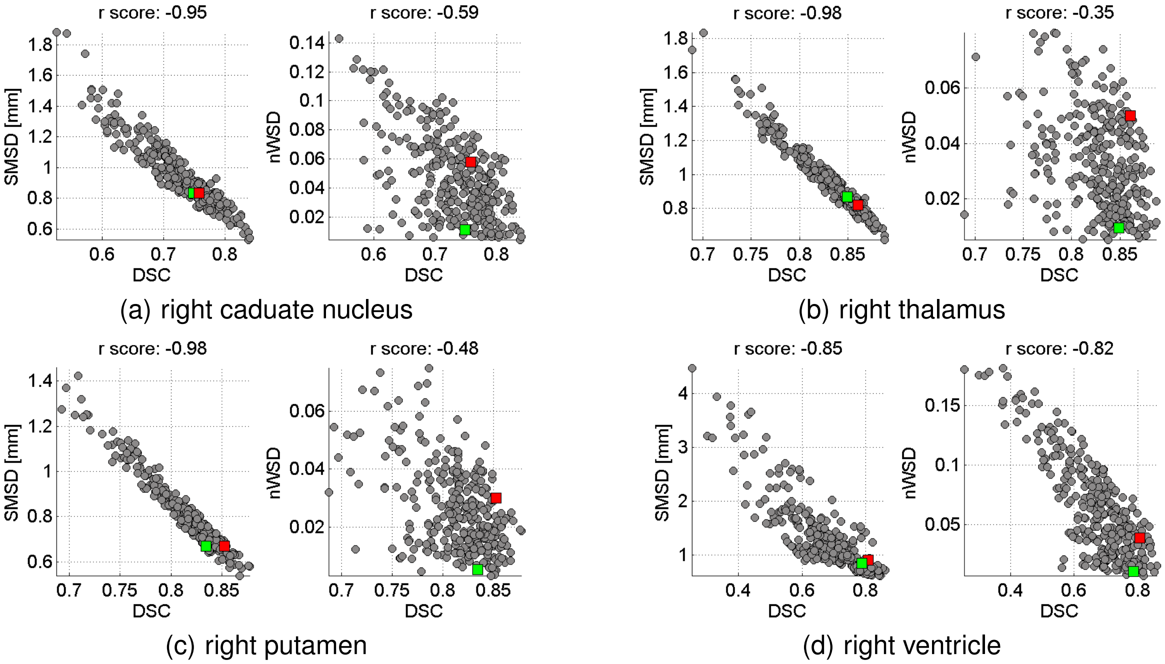
The proposed (DSC,nWSD) scoring system allows us to interpret the quality of the nonlinear registration in a much richer way than by just reporting the resulting DSC scores. For instance, for two registration problems that achieve the same DSC score we can now recognise the sources of imperfections between the aligned segmentations. A low nWSD score would hint us that the imperfection is due to pose misalignment while a high score tells us that there is a shape mismatch. Such information can help to understand better the behaviour of registration methods and the influence of parameters such as the amount of regularization.

One of the most important uses of the (DSC,nWSD) evaluation system is for comparing different registrations yielding slightly different DSC scores (or any other score mentioned in Section II). In the literature, it is common to assume that a slightly better DSC score is an indicator of a better registration (or segmentation). However, the shape variations that one obtains for exactly the same DSC score, as shown in Section II, raise some concerns on the validity of this assumption. By considering DSC only, one cannot understand whether the increase in the score is a consequence of a truly better registration or just a result of a better pose alignment while substantial shape mismatches might be present. Providing a quantification of the shape discrepancy through nWSD, removes this ambiguity. Below we demonstrate this with some visual examples. We examine four pairs of registrations, one pair for each structure, which yield slightly different DSC and SMSD scores.

From the graphs shown in Figure 9 we select two cases for each structure independently. These selected registrations are indicated by red and green points. For each of these registrations, we provide a visual interpretation of the nWSD score. We extract the segmentations of the aligned structures of interest and correct for remaining pose misalignments using the iterative closest point (ICP) algorithm. We determine the residual surface distances and colour-encode these on the surface meshes shown in Figure 10. Blue corresponds to lower residuals. For each structure, Registration #1 (#2) corresponds to the green (red) point in the respective graph in Figure 9. For the ease of demonstration, we only show either the target or the warped source segmentation, whichever shows higher residuals. The meshes are rendered from two different viewpoints. In the accompanying table we provide the DSC, SMSD, and nWSD scores for each structure and each registration.

Observing Figure 10 we notice that although the DSC scores for Registration #2 for each structure are higher, these registrations also show much higher discrepancy between the aligned surfaces after correcting for the pose misalignments which is reflected in the nWSD scores. It is debatable if these registrations are really better, and it would have been impossible to notice these differences by only considering DSC. nWSD detects these differences and assigns high scores to Registration #2. If the results shown in Registration #2 were truly better they would have also yielded lower nWSD scores. In fact, by equipping the segmentation-based evaluation system with nWSD, we can now define the following rule: if two registrations yield similar nWSD scores, and one of them has higher DSC, that one truly better aligns the delineations of the corresponding anatomical structures and therefore, has a better quality.

In summary, the experiments indicate that only relying on commonly used scoring functions, such as DSC, is not sufficient for discriminating between registrations. Registrations of different quality can be assigned similar scores. Confirming the conclusions of [17], we also see that alternative scores such as SMSD actually do not provide additional information for most of the structures. Furthermore, comparing different registration results (different methods or problems) based on DSC, or any other similar score, does not necessarily



(a) right caduate nucleus

(b) right thalamus

(c) right putamen

(d) right ventricle

Fig. 9: The DSC vs. SMSD and DSC vs nWSD scoring plots for the 306 registrations. In the graphs, each point corresponds to a different registration problem, i.e. different source-target image pair. The scores are computed for each registration by aligning the manual segmentations of the target and source image using the computed transformations. We compute the scores based on four subcortical structures: (a) caudate, (b) thalamus, (c) putamen and (d) ventricle. We note that in (a)-(c) DSC and SMSD are very correlated while nWSD and DSC are much less correlated. This shows that the information provided by nWSD is indeed not captured by DSC. The correlation scores (r-values) are given in the titles of each plot. In each plot we also highlight two points in red and in green, which we elaborate further in Fig. 10 and in the text.

provide a valid conclusion. nWSD, on the other hand, provides additional information that is not captured by the commonly used scores. Our experiments illustrate that jointly using DSC and nWSD achieves a much richer characterization and a higher discriminative power than either one of them alone. It provides the ability to interpret the imperfections in alignments as well as better means for comparisons.

V. Conclusion

This paper explored a new score, called normalised Weighted Spectral Distance (nWSD), for segmentation-based evaluation. We showed that commonly used measures, such as Dice's coefficient (DSC), are not discriminative enough in measuring discrepancies between two binary label maps. They cannot make the distinction between simple pose alignments and substantial shape mismatches. As a result they yield similar scores to a wide range of segmentation pairs. In order to overcome these shortcomings, we explored the use of a complementary measure, namely nWSD, which measures shape discrepancies between two binary label maps based on spectra of Laplace operator. Through different synthetic experiments we demonstrated that nWSD is able to quantify the shape differences other scores are indifferent to. Furthermore, theoretical and practical properties of nWSD make it a practical measure complementary to existing scores. We further showed that nWSD in combination with standard metrics, such as DSC, provides higher discrimination power in segmentation-based evaluation. nWSD has the potential to be an important component in segmentation-based evaluation studies that can be applied to future studies as well as to retrospective studies for re-evaluation. We will support those studies wanting to take advantage of nWSD by making our MATLAB® implementation available upon request.

References

[1] J. B. Maintz and M. A. Viergever, “A survey of medical image registration,” *Medical Image Analysis*, vol. 2, no. 1, pp. 1–36, 1998.

[2] W. R. Crum, T. Hartkens, and D. L. Hill, “Non-rigid image registration: theory and practice,” *Br J Radiol*, vol. 77 Spec No 2, pp. S140–153, 2004.

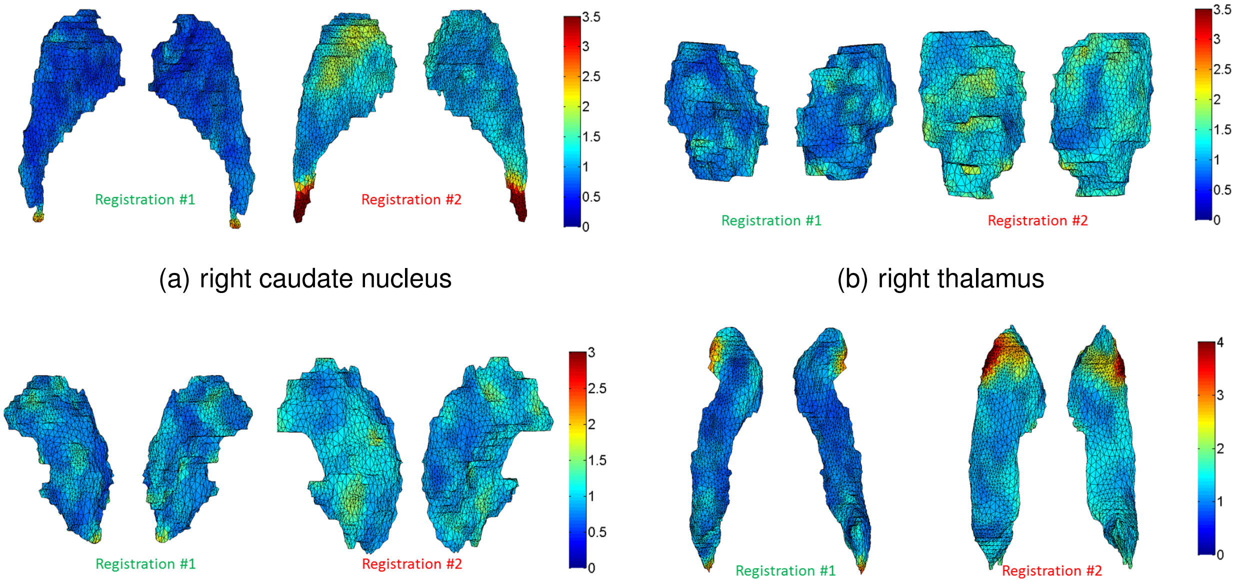
[3] A. Gholipour, N. Kehtarnavaz, R. Briggs, M. Devous, and K. Gopinath, “Brain functional localization: a survey of image registration techniques,” *IEEE Trans Med Imaging*, vol. 26, no. 4, pp. 427–451, Apr 2007.

[4] P. J. Slomka and R. P. Baum, “Multimodality image registration with software: state-of-the-art,” *Eur. J. Nucl. Med. Mol. Imaging*, vol. 36 Suppl 1, pp. 44–55, Mar 2009.

[5] Y. Zheng, C. Kambhamettu, T. Bauer, and K. Steiner, “Accurate estimation of pulmonary nodule's growth rate in ct images with nonrigid registration and precise nodule detection and segmentation,” in *Computer Vision and Pattern Recognition Workshops, 2009. CVPR Workshops 2009. IEEE Computer Society Conference on.* IEEE, 2009, pp. 101–108.

[6] P. Lorenzen, M. Prastawa, B. Davis, G. Gerig, E. Bullitt, and S. Joshi, “Multi-modal image set registration and atlas formation,” *Medical image analysis*, vol. 10, no. 3, pp. 440–451, 2006.

[7] F. L. Bookstein, ““Voxel-Based Morphometry” Should Not Be Used with Imperfectly Registered Images,” *NeuroImage*, vol. 14, no. 6, pp. 1454–1462, 2001.



(a) right caudate nucleus

(b) right thalamus

(c) right putamen

(d) right ventricle

	(a)		(b)		(c)		(d)	
	Reg. #1	Reg. #2	Reg. #1	Reg #2	Reg. #1	Reg #2	Reg. #1	Reg #2
DSC	0.75	0.76	0.85	0.86	0.83	0.85	0.79	0.81
SMSD [mm]	0.83	0.83	0.87	0.82	0.67	0.67	0.86	0.91
nWSD	0.011	0.058	0.010	0.050	0.005	0.03	0.011	0.039

Fig. 10: Figures compare the results of two different registrations for each structure. The registrations are indicated as green (#1) and red (#2) dots in the corresponding graphs in Figure 9. For each registration the segmentations of the structures of interest are extracted from the target and the warped source image. In order to give a visual interpretation of what nWSD captures, we correct for remaining errors in pose, and then determine the residual surface distance. The surface meshes display either one of the target or warped source segmentation, and we only show the one with higher residuals for ease of demonstration. The colours correspond to the local residual distances. We notice that for each structure the Registration #2 displays a substantially higher residual which is captured by nWSD. This indicates high shape mismatches between the target and the warped source segmentations. The table provides the DSC, SMSD and nWSD scores for each structure and each registration shown in the figure. Notice that for all cases Registration #2 has higher DSC despite higher shape discrepancies quantified by nWSD.

[8] T. Rohlfing, "Transformation Model and Constraints Cause Bias in Statistics on Deformation Fields," in *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2006.

[9] W. R. Crum, L. D. Griffin, D. L. G. Hill, and D. J. Hawkes, "Zen and the Art of Medical Image Registration: Correspondence, Homology, and Quality," *NeuroImage*, vol. 20, no. 3, pp. 1425–1437, 2003.

[10] J. A. Schnabel, C. Tanner, A. D. Castellano-Smith *et al.*, "Validation of Nonrigid Image Registration using Finite-Element Methods: Application to Breast MR Images," *IEEE Transactions on Medical Imaging*, vol. 22, no. 2, pp. 238–247, 2003.

[11] K. Murphy, B. Van Ginneken, J. Reinhardt, S. Kabus, K. Ding, X. Deng, K. Cao, K. Du, G. Christensen, V. Garcia *et al.*, "Evaluation of registration methods on thoracic ct: The empire10 challenge." *IEEE transactions on medical imaging*, vol. 30, no. 11, p. 1901, 2011.

[12] J. Vandemeulebroucke, D. Sarrut, and P. Clarysse, "The POPI-model, a point-validated pixel-based breathing thorax model," in *International Conference on the Use of Computers in Radiation Therapy*, 2007.

[13] S. Kabus, T. Klinder, K. Murphy, B. van Ginneken, C. Lorenz, and J. Pluim, "Evaluation of 4D-CT lung registration," in *International Conference on Medical Image Computing and Computer Assisted Intervention*, 2009.

[14] P. Hellier, C. Barillot, I. Corouge *et al.*, "Retrospective Evaluation of Intersubject Brain Registration," *IEEE Transactions on Medical Imaging*, vol. 22, no. 9, pp. 1120–1130, 2003.

[15] R. Schestowitz, C. Twining, V. Petrovic, T. Cootes, B. Crum, and C. Taylor, "Non-rigid registration assessment without ground truth," in *Medical Image Understanding and Analysis*, 2006.

[16] J. H. Song, G. E. Christensen, J. A. Hawley, Y. Wei, and J. G. Kuhl, "Evaluating Image Registration using NIREP," in *Workshop on Biomedical Image Registration*, 2010.

[17] A. Klein, J. Andersson, B. A. Ardekani *et al.*, "Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration," *NeuroImage*, vol. 46, no. 3, pp. 786–802, 2009.

[18] B. T. Yeo, M. R. Sabuncu, R. Desikan, B. Fischl, and P. Golland, "Effects of registration regularization and atlas sharpness on segmentation accuracy," *Medical Image Analysis*, vol. 12, no. 5, pp. 603–615, 2008.

[19] T. Rohlfing, "Image Similarity and Tissue Overlaps as Surrogates for Image Registration Accuracy: Widely Used but Unreliable," *IEEE Transactions on Medical Imaging*, Aug 2011.

[20] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[21] E. Konukoglu, B. Glocker, A. Criminisi, and K. M. Pohl, "WESD - weighted spectral distance for measuring shape dissimilarity," *pre-print*, 2012, arXiv:1208.5016v1 [cs.CV].

[22] W. R. Crum, O. Camara, and D. Hill, "Generalized Overlap Measures for Evaluation and Validation in Medical Image Analysis," *IEEE Transactions on Medical Imaging*, vol. 25, no. 11, pp. 1451–1461, 2006.

[23] R. Courant and D. Hilbert, *Method of Mathematical Physics, vol I.* Interscience Publishers, 1966.

[24] M. Reuter, F.-E. Wolter, and N. Peinecke, "Laplace-Beltrami spectra as 'Shape-DNA' of surfaces and solids," *Computer-Aided Design*, vol. 38, pp. 342–66, 2006.

[25] M. H. Protter, "Can one hear the shape of a drum? Revisited," *SIAM Review*, vol. 29, no. 2, pp. 185 – 197, June 1987.

[26] D. V. Vassilevich, “Heat kernel expansion: user’s manual,” *Physics Reports*, vol. 388, no. 5-6, pp. 279–360, 2003.

[27] H. Weyl, “Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen,” *Math Ann*, pp. 441–69, 1912.

[28] M. Kac, “Can one hear the shape of a drum?” *The American Mathematical Monthly*, vol. 73, no. 7, pp. 1–23, 1966.

[29] J. Sun, M. Ovsjanikov, and L. Guibas, “A concise and provably informative multi-scale signature based on heat diffusion,” in *Eurographics Symposium on Geometry Processing*, 2009.

[30] H. P. McKean and I. M. Singer, “Curvature and the Eigenvalues of the Laplacian,” *Journal of Differential Geometry*, vol. 1, pp. 43–69, 1967.

[31] L. Smith, “The asymptotics of the heat equation for a boundary value problem,” *Inventiones Mathematicae*, vol. 63, pp. 467–493, 1981.

[32] C. Gordon, D. Webb, and S. Wolpert, “Isospectral plane domains and surfaces via riemannian orbifolds,” *Inventiones Mathematicae*, vol. 110, no. 1, pp. 1–22, 1992.

[33] C. Gordon, P. Perry, and D. Schueth, “Isospectral and isoscattering manifolds: a survey of techniques and examples,” *Contemporary Mathematics*, no. 387, pp. 157–180, 2005.

[34] M. Niethammer, M. Reuter, F.-E. Wolter, S. Bouix, N. Peinecke, M.-S. Koo, and M. E. Shenton, “Global medical shape analysis using the Laplace-Beltrami spectrum,” in *Medical Image Computing and Computer Assisted Intervention*, 2007.

[35] M. Reuter, F.-E. Wolter, M. Shenton, and M. Niethammer, “Laplace-beltrami eigenvalues and topological features of eigenfunctions for statistical shape analysis,” *Computer-Aided Design*, vol. 41, pp. 739–55, 2009.

[36] F. Mémoli, “A spectral notion of gromov-wasserstein distance and related methods,” *Applied and Computational Harmonic Analysis*, vol. 30, no. 3, pp. 363–401, 2010.

[37] E. T. Whittaker and G. N. Watson, *A course of modern analysis*. Cambridge Mathematical Library, 1996.

[38] W. F. Ames, *Numerical Methods for Partial Differential Equations*. Academic Press, 1977.

[39] W. E. Arnoldi, “The principle of minimized iterations in the solution of the matrix eigenvalue problem,” *Quarterly of Applied Mathematics*, vol. 9, no. 3, pp. 17–29, 1951.

[40] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache, “Diffeomorphic demons: Efficient non-parametric image registration,” *NeuroImage*, vol. 45, no. 1, pp. S61–S72, 2009.