Patch-based Evaluation of Image Segmentation

Christian Ledig* Wenzhe Shi Wenjia Bai Daniel Rueckert
Department of Computing, Imperial College London
180 Queen's Gate, London SW7 2AZ, UK
christian.ledig@imperial.ac.uk

Abstract

The quantification of similarity between image segmentations is a complex yet important task. The ideal similarity measure should be unbiased to segmentations of different volume and complexity, and be able to quantify and visualise segmentation bias. Similarity measures based on overlap, e.g. *Dice score, or surface distances,* e.g. *Hausdorff distance, clearly do not satisfy all of these properties. To address this problem, we introduce* ***P****atch-based* ***E****valuation of* ***I****mage* ***S****egmentation (PEIS), a general method to assess segmentation quality. Our method is based on finding patch correspondences and the associated patch displacements, which allow the estimation of segmentation bias. We quantify both the agreement of the segmentation boundary and the conservation of the segmentation shape. We further assess the segmentation complexity within patches to weight the contribution of local segmentation similarity to the global score. We evaluate PEIS on both synthetic data and two medical imaging datasets. On synthetic segmentations of different shapes, we provide evidence that PEIS, in comparison to the Dice score, produces more comparable scores, has increased sensitivity and estimates segmentation bias accurately. On cardiac magnetic resonance (MR) images, we demonstrate that PEIS can evaluate the performance of a segmentation method independent of the size or complexity of the segmentation under consideration. On brain MR images, we compare five different automatic hippocampus segmentation techniques using PEIS. Finally, we visualise the segmentation bias on a selection of the cases.*

1. Introduction

The validation of automatic segmentation methods usually relies on the comparison to reference segmentations, ideally annotated by an expert observer. However, evaluation is inherently difficult as similarity features are hard to define and potentially subjective. Furthermore manual expert segmentations cannot be considered as gold standard as they are subject to both significant inter- and intra-subject variability [4, 9, 23]. Warfield *et al*. [23] presented a popular method, STAPLE, to estimate a probable ground truth given several manual expert segmentations. However, assuming a reliable ground truth or reference segmentation is available, the problem of comparison remains. The importance of quantifying segmentation accuracy in the context of biological imaging was formulated almost 40 years ago [20, 24]. However, while the development of segmentation algorithms is an active research area, the progress in developing evaluation methods is more limited [5, 20, 26].

The measurement of similarity between two different sets or image segmentations is thus an ongoing challenge [3, 8, 19, 21]. The most popular approaches to quantify segmentation similarity in medical imaging are based on overlap [7, 8], with the Dice coefficient being the most renowned representative. With the so called *ratio model* Tversky [19] presented a framework that generalised the widely used Dice [8] and Jaccard [12] similarity index. A thorough examination of the Dice and Jaccard index is presented in [5]. Generalised overlap measures focusing on fuzzy multi-label segmentations are described in Crum *et al*. [7]. In [3], the authors modified the ratio model by weighting wrongly segmented pixels in 2D, or voxels in 3D respectively, based on their spatial distance from the reference label. Another family of measures is based on the surface distance between segmentations [2, 3, 4, 10, 18]. Examples are the Yasnoff discrepancy [24] or the Hausdorff distance [11], which is highly sensitive to errors [18]. Taking the distance between misclassified voxels and ground truth into account is important [3, 7, 18].

Next to this, other methods have been proposed often with focus on a particular application. Juneja *et al*. [13] introduced the validation index which is based on multiple expert segmentations to support radiotherapy planning. In [18] the authors propose to use statistics based on surface distances to estimate segmentation bias. Recently a method based on genetic programming was proposed [22] that com-

*This work is partially funded under the Seventh Framework Programme (FP7) by the European Commission.

bines single measures for colour image segmentation.

A good overview over different segmentation evaluation approaches is given in [20, 25, 26].

1.1. Contribution and Overview

We postulate desired attributes of a similarity measure for image segmentation inspired by [21]:

- *Comparable scores:* Scores calculated on segmentations of varying size and boundary complexity should be comparable. This is desirable as it renders the comparison of automatic segmentation methods less sensitive towards the evaluation dataset or characteristic of the reference segmentation under consideration.

- *Shape conservation:* The measure should assess the preservation of the shape of segmentation boundaries. This is desirable as varying definitions of segmentation protocols or the presence of partial volume effects in low-resolution images can produce different segmentation boundaries that are, however, similar in shape.

- *Boundary conformity:* It should measure the agreement of segmentation boundaries. This is desirable as a good segmentation not only preserves shape but also matches the segmentation boundaries.

- *Segmentation bias:* The measure should allow the detection and visualisation of segmentation bias, such as over- or under-segmentation. This is desirable as it provides an insight into the differences between segmentations. Visualisation of segmentation bias can also support the development and tuning of automatic segmentation methods.

- *Similarity:* The ability to quantify *visual* similarity rather than pure overlap. This is desirable as especially segmentations of structures with a high surface to volume ratio, such as vessel segmentations, can be highly similar while having imperfect overlap.

Based on the ratio model [19], we propose *Patch-based Evaluation of Image Segmentation* (PEIS). We formulate PEIS for binary segmentations. While PEIS generalises to multi-label segmentations, this is beyond the scope of this manuscript and left for future work.

For each voxel we find a patch displacement vector that locally transforms the reference segmentation into the test segmentation. This voxel-wise displacement allows the estimation and visualisation of segmentation bias. It furthermore allows the quantification of how well a test segmentation conserves shape and matches segmentation boundaries. Our approach is inspired by patch-based methods that have been used successfully in image synthesis [16] or image denoising [14]. More recently patch-based approaches have also enjoyed increasing popularity in the medical imaging community applied for image segmentation [1, 6]. Secondly, we propose to employ the local segmentation complexity within patches to control the contribution of local segmentation similarity to the global score. This allows improving sensitivity in regions that are non-trivial to segment. Our approach extends pure overlap measures and allows for a quantitative assessment of segmentation bias. We evaluate PEIS on both synthetic and real medical imaging data: We compare PEIS to the Dice score on synthetic data and to both Dice score and the average surface distance (ASD), as used in [10], on real datasets from cardiac and brain magnetic resonance (MR) images. We provide visual evidence that PEIS allows the visualisation of segmentation bias.

The remaining paper is organised as follows: We will recap the original formulation of the ratio model and present PEIS in Section 2. We will then evaluate PEIS on synthetic and real datasets in Section 3. We will discuss our findings in Section 4 and formulate conclusions and outline future work in Section 5.

2. Method

2.1. Framework: The ratio model (RM)

Given a binary reference segmentation R, we are interested in measuring the similarity of a test segmentation $S = \{s_{i=1\ldots N}\}$ to $R = \{r_{i=1\ldots N}\}$. In the following we assume that $s_i, r_i \in \{0,1\}, \forall i \in \Omega \subset \mathbb{N}^d$. Here $d = 2,3$ is the image dimension and i serially numbers the N voxels in Ω. Ω is the subset of $\mathbb{N}^d$ where not both S and R are zero.
The ratio model (RM) framework, originally described by Tversky [19], can be stated as:

$$RM(S,R) = \frac{\theta f(S \cap R)}{\theta f(S \cap R) + \alpha f(S \setminus R) + \beta f(R \setminus S)} \quad (1)$$

We use a similar notation as in [19], where the scale f describes a family of similarity measures based on the global parameters θ, α, β. In the standard formulation $f(S)$ is the cardinality of the set S. For $\theta = 1$ and $\alpha = \beta = \frac{1}{2}$ this model reduces to the widely used Dice coefficient [8]. For $\theta = 1$ and $\alpha = \beta = 1$ the formulation yields the Jaccard coefficient respectively [12]. More details on the characteristics of this formulation can be found in Tversky [19].

In the context of classification, we can consider $f(S \cap R)$ as true positive (TP), $f(S \setminus R)$ as false positive (FP) and $f(R \setminus S)$ as false negative (FN) fraction of the test segmentation S and the reference segmentation R. We summarise the false positives and false negatives as false labels ($\mathrm{FL} = \mathrm{FP} + \mathrm{FN}$). We thus model the non-directional case, $\alpha = \beta$, where the denominator reduces to $\theta\mathrm{TP} + \alpha\mathrm{FL}$. We equally penalise false positives and false negatives. In the following we present a novel way of modelling the scale f.

2.2. Patch-based ratio model (PBRM)

Standard overlap measures are restricted to the evaluation of the voxel-wise agreement of segmentations. We propose a novel patch-based approach to relax this constraint by finding a voxel-wise displacement that transforms the reference segmentation R *locally* into the test segmentation S. We introduce $\mathcal{N}_p(\cdot)$ as a patch extraction operator. $\mathcal{N}_p(R(i))$ extracts a patch of the reference segmentation at position i, $\mathcal{N}_p(S(j))$ a patch of the test segmentation at position j accordingly.

We compare two patches by calculating the sum of absolute differences between $\mathcal{N}_p(R(i))$ and $\mathcal{N}_p(S(j))$. We define this difference as:

$$\mathcal{D}(i,j) = |\mathcal{N}_p(R(i)) - \mathcal{N}_p(S(j))| \tag{2}$$

At a certain position i we derive an optimal patch displacement $\Delta(i) = (\Delta(i)_x, \Delta(i)_y, \Delta(i)_z)$ that locally minimises $\mathcal{D}(i,j)$. This is equivalent to maximising the agreement of a fixed patch in R, $\mathcal{N}_p(R(i))$, with a moving test patch in S, $\mathcal{N}_p(S(j))$. The optimisation is described in Section 2.3. In the following, i' denotes the index that yields the best locally matching test patch $\mathcal{N}_p(S(i'))$ for a given reference patch $\mathcal{N}_p(R(i))$. Thus i' corresponds to i displaced by the optimal displacement $\Delta(i)$. For a certain displacement $\Delta(i)$, we define the spatial patch overlap between the reference patch and the displaced test patch as $\mathcal{A}(\Delta(i))$. Spatial overlaps, $\mathcal{A}(\Delta)$, are illustrated for different displacements, Δ, in yellow in Figure 1. $\mathcal{A}(\Delta)$ only depends on Δ and is thus independent of the segmentation content within the patches.

Assuming that we have found the optimal displacement $\Delta(i)$ and thus corresponding i', we define a voxel-wise segmentation similarity index $\gamma \in [0;1]$ as $\gamma(i) = 1.0 - \frac{\mathcal{D}(i,i')}{|\mathcal{N}_p(\cdot)|}$ and a spatial similarity index $\tau \in [0;1]$ as $\tau(i) = \frac{\mathcal{A}(\Delta(i))}{|\mathcal{N}_p(\cdot)|}$. These indices are normalised by the number of voxels within a patch, $|\mathcal{N}_p(\cdot)|$, and thus the maximal possible patch dissimilarity. By defining τ based on the amount of spatial overlap of the corresponding patches, we ensure that both γ and τ take on comparable values. Our formulation thus allows the averaging of γ and τ to a combined similarity score $\eta(i) = 0.5 \times (\gamma(i) + \tau(i))$.

In the notation of the ratio model [19], $\sum_i \eta(i)$ denotes the TP fraction, $f(S \cap R)$, while $\sum_i (1.0 - \eta(i))$ summarises the falsely segmented fraction, $f(S \setminus R)$ and $f(R \setminus S)$.

We formulate a patch-based ratio model (PBRM) as:

$$PBRM(S,R) = \frac{\theta \sum_i \eta(i)}{\theta \sum_i \eta(i) + \alpha \sum_i (1.0 - \eta(i))} \tag{3}$$

2.3. Optimisation of local patch correspondences

We ensure that we find a *locally* optimal patch correspondence, i', by assuming a monotonous decreasing

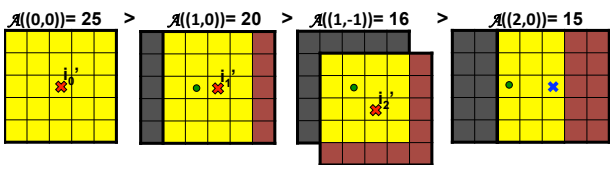

Figure 1: Decreasing *spatial* overlap $\mathcal{A}$ (yellow) in 2D between the *supports* of a reference patch (grey) and a test patch (dark red) for increasing patch displacements $\Delta = (\Delta_x, \Delta_y)$. If two displacements have the same L1-norm (both examples shown on the right) our method favours a diagonal displacement. The reference centre is indicated as circle. The centres of the test patches are indicated as crosses and correspond to displacements shown in Figure 2.

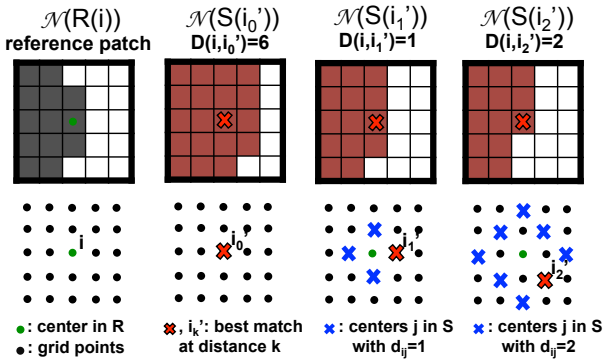

Figure 2: Calculation of optimal patch centers i'_k at different distance-levels k. An exemplary reference patch is shown in grey, while the best matching patches at the three corresponding distance-levels ($k = 0, 1, 2$) are shown in dark red.

$\mathcal{D}(i,j)$ with increasing distance d_{ij} from the reference position i. Here d_{ij} denotes the L1-distance between the position of two voxels i and j in image space. To obtain i', we thus firstly calculate solutions i'_k for different levels of distance k that minimise $\mathcal{D}(i,j)$ for all j with L1-distance d_{ij} equal to k ($\forall j : d_{ij} = k$). At a certain level k we only consider test patches centred at a voxel j with L1-distance k from i, $d_{ij} = k$. Specifically, at a certain level k we solve the following optimisation problem to calculate i'_k for a given position i in the reference segmentation:

$$\begin{aligned} i'_k = & \underset{j}{\operatorname{argmin}} \quad \mathcal{D}(i,j) \\ & \text{s. t.} \qquad d_{ij} = k \end{aligned} \tag{4}$$

The optimisation process for an exemplary reference patch is shown in Figure 2. In the case that there are multiple possible solutions that minimise $\mathcal{D}(i,\cdot)$, we choose the first found displacement that yields the maximal spatial patch

overlap, $\mathcal{A}(\Delta)$, which is illustrated in Figure 1. This will, for a certain distance-level k, favour diagonal over purely horizontal or vertical displacements.

We then calculate the *locally* optimal voxel-specific displacement $\Delta(i)$, determined by i', by solving the following optimisation problem:

$$\begin{aligned} i' = & \underset{i'_k}{\text{argmin}} \quad \mathcal{D}(i, i'_k) \\ & \text{s. t.} \qquad \mathcal{D}(i, i'_0) \geq \mathcal{D}(i, i'_1) \geq \cdots > \mathcal{D}(i, i'_k) \end{aligned} \tag{5}$$

The monotonicity constraint ensures to find the best matching patch locally. The proposed strategy also renders the search very efficient, as it provides an efficient stopping criterion for the calculation of i'_k. The calculation can be stopped once $\mathcal{D}(i, i'_{k+1}) > \mathcal{D}(i, i'_k)$. With this formulation we avoid setting a fixed search neighborhood. This is usually required to allow for a tractable computation. Hence, our approach can find arbitrary large displacements.

2.4. Patch-based Evaluation of Image Segmentation (PEIS)

Determining the parameters θ, α globally yields a measure to which correctly or falsely segmented voxels equally contribute. For large structures with a low surface to volume ratio this is problematic as the overall measure is dominated by the massive number of true positives in the interior of the structure. We address this problem and introduce spatially varying parameters $\theta(i), \alpha(i)$ into the model. The following formulation presents a data-driven way to increase the measure's sensitivity in the relevant regions, *e.g.* regions close to the boundary, when comparing structures.

We now describe the following complete patch-based spatially varying ratio model:

$$\text{PEIS}(S,R) = \frac{\sum_{\mathbf{i}} \theta(i)\eta(i)}{\sum_{\mathbf{i}} \theta(i)\eta(i) + \sum_{\mathbf{i}} \alpha(i)(1-\eta(i))} \tag{6}$$

There are different ways to model $\theta(i)$ and $\alpha(i)$. We choose $\theta(i)$ and $\alpha(i)$ based on the complexity of the reference patch under consideration. We quantify this complexity by counting the number of facets, $n_{\text{facets}}(i)$, that separate adjacent voxels with different labels in the reference patch, $\mathcal{N}_p(R(i))$, with centre i. We then calculate $\theta(i)$ and $\alpha(i)$ as:

$$\theta(i) = \frac{n_{\text{facets}}(i)}{n^{\max}_{\text{facets}}} \qquad \alpha(i) = 1.0 - \theta(i) \tag{7}$$

Here $n^{\max}_{\text{facets}}$ is chosen in the order of the number of voxels within a patch of maximal complexity. For a given patch width or patch size, p_w, we choose $n^{\max}_{\text{facets}} = 4 \times (p_w - 1)$ (2D) or $n^{\max}_{\text{facets}} = 4 \times (p_w - 1) \times p_w$ (3D). This corresponds to the number of facets contained in a patch separated by a diagonal line (2D) or diagonal plane (3D). We threshold $n_{\text{facets}}(i)$ at $n^{\max}_{\text{facets}}$ to ensure $0 \leq \theta(i) \leq 1$.

With this formulation the contribution of correct labels, $\theta(i)$, is higher in patches that are difficult to segment (high complexity) than in those that have little complexity or are even homogeneous ($n_{\text{facets}}(i)$=0). In contrary, false labels of "easy to segment" or quite homogeneous regions have a high weight, $\alpha(i)$, and thus degrade the measure more seriously than false labels in "difficult to segment" patches.

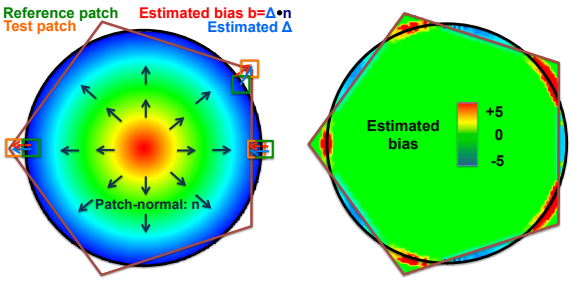

Figure 3: left: Example comparison of a polygon (test segmentation, outlined in dark red) to a circle (reference, outlined in black). Patch-correspondences of fixed reference patches (green) and moving test patches (orange). Vectors represent patch-normals $\mathbf{n}$ (dark blue), optimal patch displacements Δ (light blue) and the projections of Δ on $\mathbf{n}$ (red) which represents the estimated segmentation bias. The distance transform of the reference segmentation is shown in a rainbow colour scheme. right: Estimated voxel-wise segmentation bias $b(i) = \mathbf{n}(i) \cdot \Delta(i)$ within the domain Ω.

2.5. Quantification of Segmentation Bias

In addition to the similarity score, the presented approach also provides a displacement vector $\Delta(i) = (\Delta(i)_x, \Delta(i)_y, \Delta(i)_z)$ quantifying the local spatial difference between test and reference segmentation. In the following, we derive a single measure quantifying the segmentation bias based on Δ. Specifically, we are interested if a segmentation is too generous ("over-segmentation") or too strict ("under-segmentation").

First we calculate for a patch with centre i a vector that we call patch-normal $\mathbf{n}(i)$. We calculate $\mathbf{n}(i)$ based on the Euclidean distance transform, $\mathcal{T}(R)$, of the zero level-set in the reference segmentation. Specifically we compute $\mathbf{n}(i)$ as the average first derivative of $\mathcal{T}(R)$ at segmented voxels in the reference patch $\mathcal{N}_p(R(i))$. This can be formalised as:

$$\mathbf{n}(i) = -\frac{1}{Z} \sum_{j \in \mathcal{N}_p(R(i)))} \delta(R(j)) \nabla \mathcal{T}(R)(j) \tag{8}$$

Here j indexes the voxels in the reference patch $\mathcal{N}_p(R(i)))$

and $\delta(\cdot)$ indicates if $R(j)$ is segmented. Z is a factor that normalises $\mathbf{n}(i)$ to unit length. This defines $\mathbf{n}(i)$ as a patch-normal representing the most significant direction of the transition "from the segmented to the unsegmented region".

Together with the calculated patch displacement $\Delta(i)$, the patch-normal $\mathbf{n}(i)$ allows the quantification of over- or under-segmentation by calculating the scalar product $b(i) = \mathbf{n}(i) \cdot \Delta(i)$. If a certain patch is displaced along $\mathbf{n}(i)$ then $b(i) > 0$ indicates over-segmentation. $b(i) < 0$ indicates under-segmentation accordingly. Figure 3 illustrates the bias estimation based on the example of comparing a five sided regular polygon to a reference circle.

We further calculate the weighted mean μ_b and the standard deviation σ_b to quantify systematic segmentation bias. As weights we employ the $\theta(i)$ to focus the bias calculation on patches that contain edges. The translational bias is estimated for all directions, x, y, z, based on the components of $\Delta(i)$ and patch complexities $(\theta(i)_x, \theta(i)_y, \theta(i)_z)$ calculated based on facets perpendicular to the corresponding directions. Tab. 1 summarises the proposed quantities.

	Calculation	Description
PEIS	Eq. 6	similarity
μ_b	$\frac{1}{\sum\theta(i)} \sum_i \theta(i)(\mathbf{n}(i) \cdot \Delta(i))$	segment. bias
σ_b	$[\frac{1}{\sum\theta(i)} \sum_i \theta(i)(b(i) - \mu_b)^2]^{\frac{1}{2}}$	shape conserv.
$\mu_{d\in\{x,y,z\}}$	$\frac{1}{\sum\theta(i)_d} \sum_i \theta(i)_d \Delta(i)_d$	trans. bias (μ)
$\sigma_{d\in\{x,y,z\}}$	$[\frac{1}{\sum\theta(i)_d} \sum_i \theta(i)_d (\Delta(i)_d - \mu_d)^2]^{\frac{1}{2}}$	trans. bias (σ)

Table 1: Overview over all quantities available through PEIS and their description.

3. Experimental Results

3.1. Data

We evaluated the presented measure PEIS on both synthetic and real data. We synthetically created geometric objects to show that PEIS yields, in contrast to the Dice score (DICE), *comparable scores* on structures of different size and has the potential to quantify *shape conservation*.

- *Circles:* We compared two-dimensional circles of varying radiuses ($C^S_{r_S}$) to two reference circles with fixed radius ($C^R_{r_R}$, $r_R = 15$ or $r_R = 80$). In each experiment the radiuses of the test segmentations, r_S, varied from $r_R - 10$ to $r_R + 10$.

- *Polygons:* We compared two-dimensional regular polygons with a varying number of sides to a reference circle with fixed radius ($r_R = 50$). The polygons approximate the circle and have a fixed volume of $V_p = \pi r_R^2$ and thus the same volume as the circle. The number of polygon sides varied from 3 to 50.

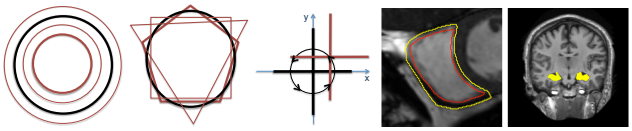

Figure 4: Overview over employed datasets. Synthetic datasets with reference segmentation outline (black) and exemplary test segmentation outlines (dark red). Medical imaging datasets with overlaid ground truth reference segmentation. Cardiac MR image with outlined right ventricle myocardium (yellow) and blood pool (red), and a MR brain scan with overlaid hippocampus segmentation (yellow).

To furthermore evaluate PEIS on vessel-like structures with high surface to volume ratio we created synthetic *crosses*. On this dataset we confirmed that PEIS accurately estimates translational bias and compares shape similarity rather than pure overlap.

- *Crosses:* We compared three-dimensional crosses with varying centre $(c_x = 10cos(t), c_y = 10sin(t), c_z = 1)$ parameterised by $t \in [0, 2\pi]$ to a reference cross centred in the origin. All crosses are one voxel thick and have a radius of 50.

We also investigated the characteristics of PEIS on real medical imaging data. We compared automatic segmentations to available manual ground truth segmentations. On a *cardiac* dataset [17] we showed that PEIS provides *comparable scores* between segmentations of different volume and topology. On a *brain* imaging dataset [15] we showed that PEIS can be used to rank automatic segmentation methods and allows the visualisation of *segmentation bias*. The employed datasets from the medical imaging domain are:

- *Cardiac data:* Automatic segmentations were calculated based on 16 training subjects used in a recent RV (right ventricle) Segmentation Challenge [17]. We calculated the Dice score, average symmetric surface distance (ASD) and PEIS score for the myocardium and blood pool as well as their union.

- *Brain data:* Automatic segmentations were calculated on 20 subjects used for evaluation in a recent whole-brain segmentation challenge [15]. We compared DICE and PEIS on the hippocampus segmentations calculated by the five best performing methods in the challenge. We focused on the hippocampus as its accurate segmentation has been proven to be of high value for the diagnosis and outcome prediction of dementia patients, in Alzheimer's Disease in particular.

An overview over the employed data is provided in Figure 4.

CVPR #945

3.2. Synthetic Data

On synthetically generated data, we observed that PEIS is in contrast to the Dice coefficient fulfilling the claimed characteristics: *comparable scores, shape conservation, boundary conformity, segmentation bias* and *similarity*.

In a first experiment we used the synthetically generated *circles*. This experiment, as summarised in Figure 5, confirmed that the measured segmentation bias μ_b quantified *boundary conformity* as it represented the true bias for both under- and over-segmentation. The low standard deviation σ_b underlined *shape conservation*. PEIS provided *comparable scores* for different reference radiuses. In contrast to that, Dice scores increased with increasing radiuses. DICE was reasonably discriminative only for deviations from a circle with a rather small radius (C_{15}^R), as it tended to degenerate for circles with larger radiuses (C_{80}^R). The vast number of true positives in the interior of large structures dominated the overlap measure and DICE lost its sensitivity to small deviations from the reference.

In a second experiment, we compared the *polygons* to a reference circle of fixed radius. Figure 6 illustrates the calculated similarities. With high Dice scores, *e.g.* 0.91 for the comparison of a square to a circle, DICE was unable to detect the significant shape differences between objects of equivalent volume. PEIS detected differences over a wide spectrum and was able to quantify the shape differences between segmentations (σ_b). A visualisation of the segmentation bias is illustrated in Figure 3, confirming the potential of PEIS to quantify and visualise *segmentation bias*.

In the final experiment on synthetic data we evaluated the ability to detect translational bias in the segmentation: For this, we translated three-dimensional *crosses* along a circular trajectory. Figure 7: While a DICE value of almost zero, due to neglectable overlap, failed to quantify any similarity, non-zero PEIS scores of 0.21±0.06 (mean±SD) confirmed that PEIS quantifies similarity rather than pure overlap. PEIS yielded *comparable scores* while moving the test cross in a constant distance around the reference cross. PEIS also provided a good estimate for the translational bias (μ_x, μ_y, μ_z) for all coordinate directions with almost zero variance ($\sigma_x = \sigma_y = \sigma_z \approx 0$).

3.3. Medical Imaging Data

In this section, we employed PEIS to evaluate segmentations of medical imaging data. Firstly, we obtained evidence that PEIS provides *comparable scores* on *cardiac data*: While both DICE and also the average symmetric surface distance [10] provided scores of significantly different levels for the myocardium and blood pool segmentation, PEIS remained comparable, as illustrated in Figure 8. This is important as it indicated that the proposed measure is indeed independent of the topology and volume of the segmentation under consideration. This result also suggested

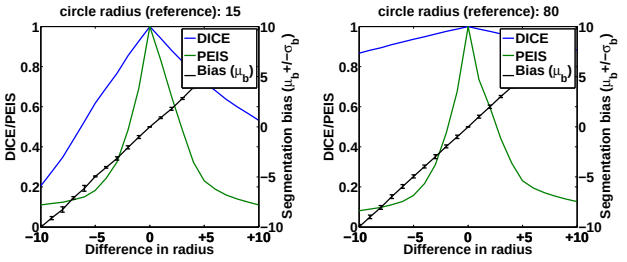

Figure 5: *Circles* experiments for reference radiuses $r_R = 15$ (left) and $r_R = 80$ (right): Different similarity scores (colours) and segmentation bias μ_b (black, σ_b as error bars).

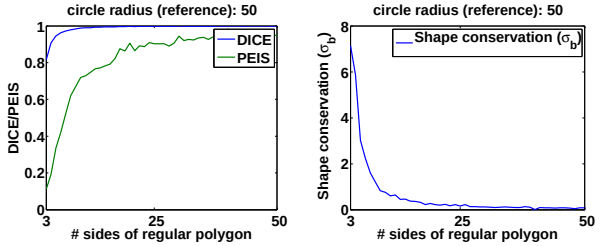

Figure 6: *Polygons* experiment: Different similarity scores (left) and estimated shape conservation, σ_b (right).

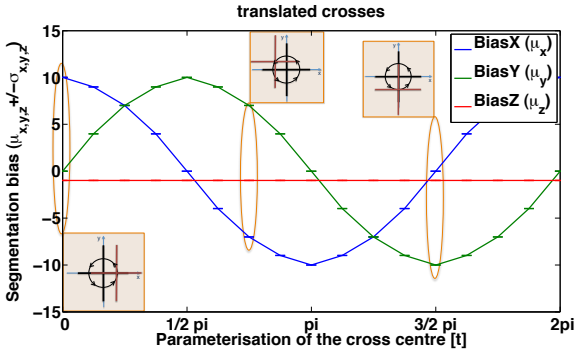

Figure 7: *Crosses* experiment: Estimated translational bias for all coordinate directions ($\sigma_{x,y,z}$ as error bars). Crosses were translated by $(c_x = 10cos(t), c_y = 10sin(t), c_z = 1$ with $t \in [0, 2\pi])$.

that PEIS has the potential to compare different segmentation methods over different datasets.

Secondly, we applied PEIS to evaluate the hippocampus segmentations of the five best performing methods in a recent whole-brain segmentation challenge [15]. The results are summarised in Table 2. Here, PEIS produced a similar ranking as DICE, which was used in the challenge. More interesting, however, was the additional information pro-

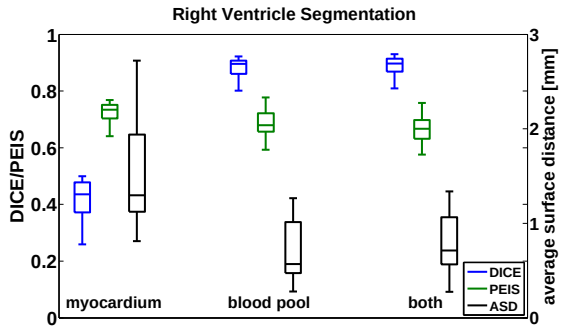

Figure 8: Similarity scores calculated on *cardiac data* segmented into myocardium and blood pool. Both DICE and the average symmetric surface distance (ASD) do not yield *comparable scores* for structures of different complexities.

vided through PEIS. The best performing method is also the method that best conserved the segmentation shape, lowest σ_b. We also noted the negative bias, μ_b, of MALP_EM which indicated a consistent under-segmentation of the hippocampus. Furthermore we were able to obtain the visualisation of the segmentation bias on right hippocampus segmentations of the five methods (Figure 9).

These experiments on real data confirmed that PEIS allows the quantification and visualisation of *segmentation bias*, yields *comparable scores* and adds information regarding *shape conservation*.

	CIS_JHU	MALP_EM	PICSL_Joint	NLS	PICSL_BC
DICE	0.851	0.861	0.862	0.866	0.870
PEIS	0.788	0.801	0.799	0.809	0.810
μ_b	-0.007	-0.069	0.011	-0.007	-0.017
σ_b	0.541	0.483	0.492	0.471	0.468

Table 2: Similarity scores (DICE, PEIS) and mean segmentation bias (μ_b) and shape conservation (σ_b) averaged over 20 hippocampus segmentations for different methods.

4. Discussion

We have shown that PEIS yields *comparable scores* on real medical imaging datasets, which sets it apart from many existing approaches such as DICE or surface-based distance measures. In the presented formulation, PEIS is not independent of the resolution level of a segmentation as the spatial similarity τ is linked to the voxel-wise patch displacement. E.g. in the *circles* experiments (cf. Figure 5), PEIS will yield lower scores if the data is resampled at a higher resolution. Dependent on the application this may or may not be appropriate. However, the estimated segmentation bias, if calculated in millimetres, is comparable. In general it is difficult to compare or even rank similarity measures as they usually provide complementary information. However, we have shown that PEIS provides a large spectrum of information including *shape conservation* and *bias estimation*. As PEIS finds patches of the reference segmentation in the test segmentation it is a directional measure. This contrasts it from DICE, which is non-directional. While non-directionality seems a desirable attribute it bares the risk that the measures' scores may be biased towards certain test segmentations. A further discussion of directionality can be found in [19]. The presented optimisation approach of PEIS is efficient, has no fixed search window size and ensures to find a local optimum, which is desired for this application. In the conducted experiments, we chose a patch size of $5\times5\times5$ voxels according to commonly used patch-based segmentation methods, e.g. [6]. The runtimes of PEIS and ASD were comparable and in the order of seconds (single core).

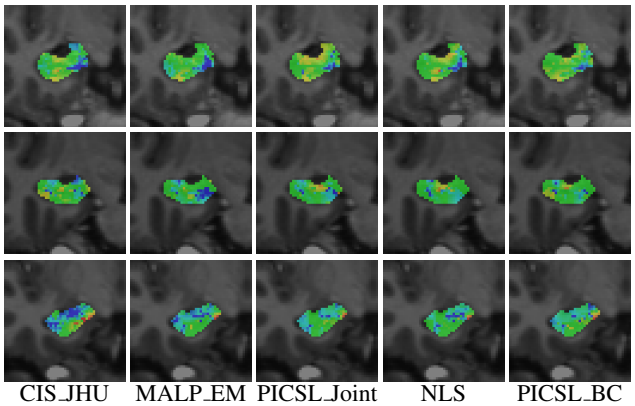

Figure 9: Visualisation in coronal view of segmentation bias for the right hippocampus of three random subjects (rows) for five automatic segmentation methods (columns). Colours relate to positive bias or over-segmentation (red) and negative bias or under-segmentation (blue).

5. Conclusion and future work

We have introduced a novel Patch-based Evaluation of Image Segmentation (PEIS) for the comparison of binary segmentations. We demonstrated on both synthetic data and real medical imaging data that PEIS has the potential to fulfil the key characteristics which are important for the evaluation of image segmentation: providing a similarity score that is both comparable over different applications and informative. The presented measure has proven to be sensitive to small, but critical, shape differences of objects with varying size. Furthermore, PEIS yields a voxel-wise bias estimate which allows the quantification of systematic bound-

ary differences of both shape and translational nature. Especially the possibility to visualise segmentation bias is valuable when comparing segmentations and can support the development and tuning of automatic segmentation methods. With the incorporation of locally varying weights based on patch complexity our approach extends the standard ratio model. This is a step from pure overlap measures towards measures with a focus on segmentation similarity.

In this work we have evaluated PEIS based on the combined quantity, $\eta = 0.5(\gamma+\tau)$. However, PEIS can be based on other scales such as the pure segmentation similarity γ or the pure spatial similarity τ. In the future it will be interesting to investigate whether PEIS can be adapted by varying the patch-size to define similarity at different levels of detail. The presented optimisation strategy can potentially be employed in other patch-based applications, such as label fusion [1, 6]. We are planning to extend PEIS to multi-label segmentations and release an implementation. The presented experiments have shown that PEIS is able to recover rigid transformations, which links PEIS to image registration. Potentially not only segmentation bias but also a deformation field relating reference and test segmentation can be recovered.

References

[1] A. J. Asman and B. A. Landman. Non-local statistical label fusion for multi-atlas segmentation. *Medical Image Analysis*, 17(2):194–208, 2013.

[2] N. Aspert, D. Santa-Cruz, and T. Ebrahimi. MESH: measuring errors between surfaces using the Hausdorff distance. In *IEEE Int. Conf. Multimedia and Expo*, pages 705–708, 2002.

[3] R. Cárdenes, R. de Luis-García, and M. Bach-Cuadra. A multidimensional segmentation evaluation for medical image data. *Computer Methods and Programs in Biomedicine*, 96(2):108–124, 2009.

[4] V. Chalana and Y. Kim. A methodology for evaluation of boundary detection algorithms on medical images. *IEEE Trans. Med. Imag.*, 16(5):642–652, 1997.

[5] H. Chang, A. H. Zhuang, D. J. Valentino, and W. Chu. Performance measure characterization for evaluating neuroimage segmentation algorithms. *NeuroImage*, 47(1):122–135, 2009.

[6] P. Coupé, J. V. Manjón, V. Fonov, J. Pruessner, M. Robles, and D. L. Collins. Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. *NeuroImage*, 54(2):940–954, 2011.

[7] W. R. Crum, O. Camara, and D. L. G. Hill. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Trans. Med. Imag.*, 25(11):1451–1461, 2006.

[8] L. R. Dice. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302, 1945.

[9] G. Gerig, M. Jomier, and M. Chakos. Valmet: A New Validation Tool for Assessing and Improving 3D Object Segmentation. pages 516–523. MICCAI, 2001.

[10] T. Heimann, B. Van Ginneken, M. A. Styner, et al. Comparison and evaluation of methods for liver segmentation from CT datasets. *IEEE Trans. Med. Imag.*, 28(8):1251–1265, 2009.

[11] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge. Comparing images using the Hausdorff distance. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993.

[12] P. Jaccard. The distribution of the flora in the alpine zone. *New Phytologist*, 11(2):37–50, 1912.

[13] P. Juneja, P. M. Evans, and E. J. Harris. The Validation Index: A new metric for validation of segmentation algorithms using two or more expert outlines with application to radiotherapy planning. *IEEE Trans. Med. Imag.*, 32(8):1481–1489, 2013.

[14] C. Kervrann and J. Boulanger. Optimal Spatial Adaptation for Patch-Based Image Denoising. *IEEE Trans. Image Processing*, 15(10):2866–2878, 2006.

[15] B. A. Landman and S. K. Warfield. MICCAI Grand Challenge and Workshop on Multi-Atlas Labeling. 2012.

[16] L. Liang, C. Liu, Y.-Q. Xu, B. Guo, and H.-Y. Shum. Real-time texture synthesis by patch-based sampling. *ACM Trans. Graphics*, 20(3):127–150, 2001.

[17] C. Petitjean, S. Ruan, D. Grosgeorge, J. Caudron, and J.-N. Dacher. Right ventricle segmentation in cardiac MRI: a MICCAI 2012 challenge. In *MICCAI Right Ventricle Segmentation Challenge Workshop*, 2012.

[18] E. Pichon, A. Tannenbaum, and R. Kikinis. A statistically based flow for image segmentation. *Medical Image Analysis*, 8(3):267–274, 2004.

[19] A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.

[20] J. K. Udupa, V. R. LeBlanc, Y. Zhuge, C. Imielinska, H. Schmidt, L. M. Currie, B. E. Hirsch, and J. Woodburn. A framework for evaluating image segmentation algorithms. *Comput. Med. Imaging Graph.*, 30(2):75–87, 2006.

[21] R. Unnikrishnan, C. Pantofaru, and M. Hebert. Toward Objective Evaluation of Image Segmentation Algorithms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 29(6):929–944, 2007.

[22] H. Vojodi, A. Fakhari, and A. M. Eftekhari Moghadam. A new evaluation measure for color image segmentation based on genetic programming approach. *Image and Vision Computing*, 31(11):877–886, 2013.

[23] S. K. Warfield, K. H. Zou, and W. M. Wells. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans. Medical Imaging*, 23(7):903–921, 2004.

[24] W. A. Yasnoff, J. K. Mui, and J. W. Bacus. Error measures for scene segmentation. *Pattern Recognition*, 9(4):217–231, 1977.

[25] H. Zhang, J. E. Fritts, and S. A. Goldman. Image segmentation evaluation: A survey of unsupervised methods. *Comput. Vis. Image Und.*, 110(2):260–280, 2008.

[26] Y. J. Zhang. A survey on evaluation methods for image segmentation. *Pattern Recognition*, 29(8):1335–1346, 1996.