Evaluating Shape Correspondence for Statistical Shape Analysis: A Benchmark Study

Brent C. Munsell, *Student Member*, *IEEE*, Pahal Dalal, *Student Member*, *IEEE*, and Song Wang, *Member*, *IEEE*

Abstract—This paper introduces a new benchmark study to evaluate the performance of landmark-based shape correspondence used for statistical shape analysis. Different from previous shape-correspondence evaluation methods, the proposed benchmark first generates a large set of synthetic shape instances by randomly sampling a given statistical shape model that defines a ground-truth shape space. We then run a test shape-correspondence algorithm on these synthetic shape instances to identify a set of corresponded landmarks. According to the identified corresponded landmarks, we construct a new statistical shape model which defines a new shape space. We finally compare this new shape space against the ground-truth shape space to determine the performance of the test shape-correspondence algorithm. In this paper, we introduce three new performance measures that are landmark independent to quantify the difference between the ground-truth and the newly derived shape spaces. By introducing a ground-truth shape space that is defined by a statistical shape model and three new landmark-independent performance measures, we believe the proposed benchmark allows for a more objective evaluation of shape correspondence than previous methods. In this paper, we focus on developing the proposed benchmark for 2D shape correspondence. However, it can easily be extended to 3D cases.

Index Terms—Statistical shape analysis, shape correspondence, point distribution model, benchmark study.

1 Introduction

In order to accurately model structural shape and its possible variation, statistical shape analysis has become a major research topic in computer vision in recent years. In statistical shape analysis, each data sample is a shape instance that is usually in the form of a smooth contour in 2D or a smooth surface in 3D. The goal is to construct a statistical shape model using these sample shape instances. Statistical shape models can be applied to address many important applications in computer vision and medical image analysis. For example, in [6], [16], statistical shape models are successfully used to guide image segmentation by detecting structures with desirable shapes. In [2], [3], statistical shape models are constructed to accurately locate the subtle differences in the corpus-callosum shapes between schizophrenia patients and normal controls.

Traditional statistics theory and tools generally handle vectors with a fixed dimension. Therefore, in statistical shape analysis, the first step is to discretize a continuous shape instance into a finite-dimension vector.[1] This is achieved by identifying a set of landmarks on each shape instance. It is critical that the landmarks identified from each shape instance be of the same number, well corresponded, and sufficiently dense to represent the underlying continuous shape instance [19]. This landmark identification step is usually referred to as *(landmark-based) shape correspondence*, which has been shown to be a major factor in determining the accuracy of the resulting statistical shape model [3], [9].

Developing more accurate and efficient shape-correspondence algorithms used in statistical shape analysis has been widely investigated in the past several years [9], [23], [24]. However, objective evaluation of the results produced by these shape-correspondence algorithms is still a very difficult problem. One major reason is the unavailability of a ground-truth shape correspondence: Given a set of real shape instances, say, kidney contours from a group of people, even the landmarks identified by different experts may be substantially different from each other [20]. Without using a ground-truth shape correspondence, Davies et al. [8], [20] introduce three general measures to describe the compactness, specificity, and generality of the statistical shape model constructed from a shape-correspondence result and suggest the use of these three measures for evaluating the shape-correspondence performance. However, as elaborated upon in Section 2.3, these three landmark-dependent measures may not objectively reflect the real shape-correspondence performance.

In this paper, we present a new benchmark study based on a given statistical shape model to achieve a more objective evaluation of shape correspondence. Specifically, the statistical shape model chosen for this paper is the widely used point distribution model (PDM) [6]. The PDM and shape correspondence have been studied in both the 2D and 3D cases. For simplicity, this paper focuses on the 2D case; however, the basic principles and algorithms outlined in this paper can be extended to 3D. For the 2D case, a PDM is defined as a $2m$-dimensional Gaussian distribution with m being the number of corresponded landmarks identified from each shape instance. In this

1. In [15], Kotcheff and Taylor introduce an algorithm to construct a statistical shape model from corresponded continuous shape instances without sampling landmarks.

• *The authors are with the Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208. E-mail: {munsell, dalalpk, songwang}@engr.sc.edu.*

Manuscript received 29 Sept. 2006; revised 27 June 2007; accepted 26 Nov. 2007; published online 18 Dec. 2007.
Recommended for acceptance by S. Carlsson.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0692-0906.
Digital Object Identifier no. 10.1109/TPAMI.2007.70841.

benchmark, we start with a given PDM by specifying a $2m$-dimensional mean-shape vector and a $2m \times 2m$ covariance matrix. By interpolating landmarks into continuous contours, this PDM defines a *(deformable or probabilistic) shape space* in which each continuous shape contour has its own probability density. We treat this shape space in the continuous domain as ground truth for our benchmark. We then randomly sample this PDM (in fact the ground-truth shape space) to generate a set of synthetic continuous shape contours. A test shape-correspondence algorithm is then applied to correspond these shape contours by identifying a new set of corresponded landmarks. Finally, we construct a new PDM from the corresponded landmarks and check whether it well describes the ground-truth shape space. This is used to evaluate the accuracy of the shape correspondence. Note that, in the proposed benchmark, the ground-truth shape space (defined by a given ground-truth PDM) may not accurately describe certain real structures from an anatomic and biomedical perspective and the shape contours used for testing shape correspondence are synthetically generated.

The remainder of this paper is organized as follows: Section 2 provides a review of the 2D PDM, shape correspondence, and Davies' three measures to evaluate shape-correspondence performance. Section 3 describes the proposed benchmark and develops three new measures to evaluate shape-correspondence performance. In Section 4, we apply the proposed benchmark and these three new measures to evaluate five available shape-correspondence algorithms. In Section 5, we discuss a possible modification to the proposed benchmark to evaluate shape correspondence by assessing its ability to construct a PDM that accurately describes the shape space of real structures. Section 6 is the conclusion.

2 PDM, Shape Correspondence, and the Evaluation of Shape Correspondence

In this section, we briefly review the principles of PDM, shape correspondence and Davies's three measures to evaluate shape-correspondence performance.

2.1 PDM and Shape Correspondence

Given n sample shape instances (or *shape contours* in the 2D case) S_a, $a = 1, 2, \ldots, n$, a PDM [6] is usually constructed in three steps: shape correspondence, shape normalization, and statistical processing.

As discussed above, shape correspondence aims at identifying corresponded landmarks from a set of continuous shape contours. More specifically, after shape correspondence, we obtain n corresponded landmark sets $\hat{V}_a$, $a = 1, 2, \ldots, n$, from S_a, $a = 1, 2, \ldots, n$, respectively. Here, $\hat{V}_a = \{\hat{\mathbf{v}}_{a1}, \hat{\mathbf{v}}_{a2}, \ldots, \hat{\mathbf{v}}_{am}\}$ are m landmarks identified from shape contour S_a and $\hat{\mathbf{v}}_{ai} = (\hat{x}_{ai}, \hat{y}_{ai})$ is the ith landmark identified along S_a. Landmark correspondence means that $\hat{\mathbf{v}}_{ai}$, $a = 1, 2, \ldots, n$, that is, the ith landmark in each shape contour, are corresponded for any $i = 1, 2, \ldots, m$.

In practice, structural shape is usually assumed to be invariant to any (uniform) scaling, rotation, and translation transformations. In shape normalization, such transformations are removed among the given n shape contours by normalizing each of the n identified corresponded landmark sets $\hat{V}_a$ to $V_a = \{\mathbf{v}_{a1}, \mathbf{v}_{a2}, \ldots, \mathbf{v}_{am}\}$, $a = 1, 2, \ldots, n$. After the shape normalization, the absolute coordinates of the corresponded landmarks, for example, $\mathbf{v}_{ai} = (x_{ai}, y_{ai})$, $a = 1, 2, \ldots, n$, are directly comparable. Procrustes analysis [13] is one of the most widely used algorithms for shape normalization.

Finally, we calculate the PDM by fitting the normalized landmarks sets $V_a = \{\mathbf{v}_{a1}, \mathbf{v}_{a2}, \ldots, \mathbf{v}_{am}\}$, $a = 1, 2, \ldots, n$, to a multivariate Gaussian distribution. Specifically, we columnize m landmarks in V_a into a $2m$-dimensional vector $\mathbf{v}_a = (x_{a1}, y_{a1}, x_{a2}, y_{a2}, \ldots, x_{am}, y_{am})^T$ and call it a *(landmark-based) shape vector* of the shape contour S_a. The mean shape vector $\bar{\mathbf{v}}$ and the covariance matrix $\mathbf{D}$ can be calculated by

$$\begin{aligned} \bar{\mathbf{v}} &= \frac{1}{n}\sum_{a=1}^{n} \mathbf{v}_a, \\ \mathbf{D} &= \frac{1}{n-1}\sum_{a=1}^{n} (\mathbf{v}_a - \bar{\mathbf{v}})(\mathbf{v}_a - \bar{\mathbf{v}})^T. \end{aligned} \tag{1}$$

The Gaussian distribution $\mathcal{N}(\bar{\mathbf{v}}, \mathbf{D})$ is the resulting PDM that attempts to model the shape space of the considered structure.

In practice, we sometimes consider only the first M ($1 \leq M \leq 2m$) eigenvectors of $\mathbf{D}$ in modeling the shape deformation. Specifically, let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{2m}$ be the eigenvalues of $\mathbf{D}$ and $\mathbf{p}_j$, $j = 1, 2, \ldots, 2m$, be the corresponding eigenvectors. The PDM that only considers the first M eigenvectors, denoted as PDM(M) for brevity, is then the multivariate Gaussian distribution $\mathcal{N}(\bar{\mathbf{v}}, \sum_{j=1}^{M} \mathbf{p}_j \lambda_j \mathbf{p}_j^T)$.

Clearly, the accuracy of the PDM is largely dependent on the performance of shape correspondence, that is, the accuracy in identifying the corresponded landmarks $\hat{V}_a$, $a = 1, 2, \ldots, n$, from the continuous shape contours S_a, $a = 1, 2, \ldots, n$. As discussed in Section 1, the performance of shape correspondence is not well defined because, in practice, a ground-truth correspondence is usually not available. Previous shape-correspondence algorithms are developed by optimizing some assumed physical or mathematical models. For example, in [2], [18], [23], 2D thin-plate splines are used to model the nonrigid deformation between different shape contours and the identified landmarks are considered to be well corresponded when a cost function based on the thin-plate bending energy is minimized. In [9], shape correspondence is considered to be of higher accuracy when it leads to a more compact PDM, that is, a PDM with smaller eigenvalues of its covariance matrix $\mathbf{D}$. In [24], shape correspondence is reduced to a medial-axis matching problem across shape contours. This method assumes that the medial axis for each shape contour has the same topology. There is also prior work on shape matching, where the corresponded landmarks are selected from two sets of points presampled from the two shape contours. For example, Belongie et al. [1] develop a shape-matching algorithm to find corresponded landmarks by matching points with similar shape context. Chui and Rangarajan [5] develop a shape-matching algorithm to find the corresponded landmarks by minimizing the thin-plate bending energy. Note that both shape-matching algorithms do not require the same number of presampled points along the two shape contours. However, the resulting corresponded landmarks along these two shape contours are required

to be the same if we want to use them for constructing statistical shape models. Because many shape-matching algorithms identify the corresponded landmarks from presampled points, they may not provide the level of accuracy required for statistical shape analysis.

2.2 Davies' Three Measures for Shape-Correspondence Evaluation

Recently, Davies et al. [8], [20] suggested the use of three general measures, *compactness*, *specificity*, and *generality*, for evaluating shape-correspondence performance in terms of the PDM construction. These three measures describe the properties of the PDM constructed from the identified landmarks and are defined as functions of the number of considered eigenvectors M, $1 \leq M \leq 2m$:

- *Compactness* evaluates shape correspondence by measuring the amount of variance of the resulting PDM. It is measured by

$$C(M) = \sum_{j=1}^{M} \lambda_j,$$

where $1 \leq M \leq 2m$. Let $C_1(M)$ and $C_2(M)$ be the compactness measures of two PDMs (PDM 1 and PDM 2, respectively) constructed from two different shape-correspondence results. If $C_1(M) \leq C_2(M)$, $\forall M \in \{1, 2, \ldots, 2m\}$, and $C_1(M) < C_2(M)$ for one or more M, we say PDM 1 is more compact than PDM 2. Therefore, the shape-correspondence result used to construct PDM 1 is better than the one used to construct PDM 2.
- *Generality* evaluates shape correspondence by measuring the resulting PDM's ability to represent unseen shape contours. In [8], [20], Davies et al. suggest the use of a "leave-one-out" cross-validation strategy to quantify the generality. Let $\mathbf{v}_a$, $a = 1, 2, \ldots, n$, be the shape vectors resulting from a shape correspondence algorithm. The generality measure $G(M)$ is calculated by the following steps: 1) Leave out one shape vector, say, $\mathbf{v}_a$, from the given n shape vectors and derive a PDM by applying (1) on the remaining $n-1$ shape vectors. We denote the resulting PDM to be PDM^a. 2) Calculate the approximation error $\epsilon_a(M)$ (based on the squared euclidean distance) using the first M eigenvectors in PDM^a to describe the left-out vector $\mathbf{v}_a$. 3) Repeat the above steps for each shape vector $\mathbf{v}_a$ in turn. The generality measure is defined as

$$G(M) = \frac{1}{n} \sum_{a=1}^{n} \epsilon_a(M).$$

Similarly to the compactness measure, let $G_1(M)$ and $G_2(M)$ be the generality measures resulting from two different shape-correspondence results. If $G_1(M) \leq G_2(M)$, $\forall M \in \{1, 2, \ldots, 2m\}$, and $G_1(M) < G_2(M)$ for one or more M, we say the first shape-correspondence result is better than the second.
- *Specificity* evaluates shape correspondence by measuring the PDM's capability to generate new "legal" shape contours. This is achieved by randomly generating a large number of shape vectors from the resulting PDM $\mathcal{N}(\bar{\mathbf{v}}, \mathbf{D})$ and then finding its nearest shape vector in $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$. The specificity measure $S(M)$ is defined as

$$S(M) = \frac{1}{N} \sum_{j=1}^{N} \|\mathbf{v}_j(M) - \mathbf{v}'_j\|^2,$$

where N is the number of randomly generated shape vectors, $\mathbf{v}_j(M)$ is a randomly generated shape vector using $\mathrm{PDM}(M)$, and $\mathbf{v}'_j$ is the shape vector in $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ that has the shortest euclidean distance to $\mathbf{v}_j(M)$. We randomly generate $\mathbf{v}_j(M)$ in the form of

$$\mathbf{v}_j(M) = \bar{\mathbf{v}} + \sum_{i=1}^{M} b_i \mathbf{p}_i,$$

where b_i, $i = 1, 2, \ldots, M$, are independently sampled from the 1D Gaussian distribution $\mathcal{N}(0, \lambda_i)$, $i = 1, 2, \ldots, M$, respectively. Similarly to the compactness measure, let $S_1(M)$ and $S_2(M)$ be the specificity measures of two PDMs (PDM 1 and PDM 2, respectively) constructed from two different shape-correspondence results. If $S_1(M) \leq S_2(M)$, $\forall M \in \{1, 2, \ldots, 2m\}$, and $S_1(M) < S_2(M)$ for one or more M, we say PDM 1 is more specific than PDM 2. Therefore, the shape-correspondence result used to construct PDM 1 is better than the one used to construct PDM 2.

2.3 Limitations of the Compactness, Generality, and Specificity Measures

In this section, we provide a simple example to demonstrate the limitations of the compactness, generality, and specificity measures introduced in Section 2.2. Let us consider an example of shape correspondence and PDM construction from five triangle shapes, as shown in Fig. 1a. Suppose we have three shape correspondence algorithms, with shape correspondence results shown in Figs. 1a, 1b, and 1c, respectively. In Fig. 1a, three landmarks are identified at the vertices of each triangle shape contour. In Figs. 1b and 1c, four landmarks are identified from each triangle shape contour. Besides the three vertices of the triangle, the shape-correspondence result in Fig. 1b also includes a landmark at the midpoint between the top vertex and the bottom-left vertex of the triangle. The shape-correspondence result in Fig. 1c includes a landmark at the midpoint between the bottom-left and bottom-right vertices of the triangle.

Using (1), we construct three different PDMs from the shape-correspondence results shown in Figs. 1a, 1b, and 1c, respectively. The dimensions of the mean shape vector and covariance matrix constructed from the landmarks shown in Fig. 1a are 6 and 6×6, respectively, whereas the dimensions of the mean shape vector and covariance matrix constructed from the landmarks shown in Fig. 1b (or Fig. 1c) are 8 and 8×8, respectively. It is easy to check that these three PDMs define exactly the same shape space in the continuous domain (by connecting landmarks sequentially with straight line segments): a triangle shape with two fixed vertices $(-2, 0)$ and $(2, 0)$ and the third vertex independently moving horizontally and vertically around the center $(0, 2)$. Both the horizontal and vertical moving distances conform to the Gaussian distribution $\mathcal{N}(0, 0.5)$. Therefore,

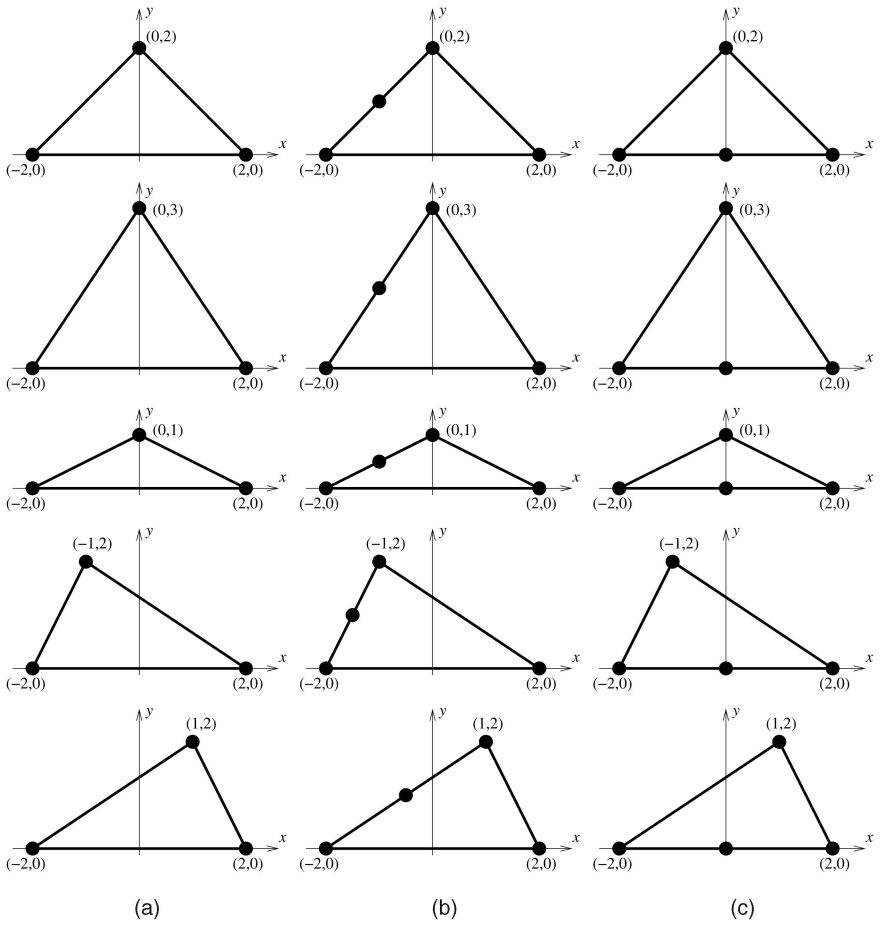


Fig. 1. An example for illustrating the limitations of the compactness, generality, and specificity measures.

the shape-correspondence results shown in Figs. 1a, 1b, and 1c should have the same accuracy.

This example shows that, for a given set of continuous shape contours, shape correspondence may not have a unique solution. Different shape-correspondence results may lead to different PDMs that describe the same shape space. As shown in Figs. 1a and 1b, one shape-correspondence algorithm may find m_1 landmarks from each shape contour and construct a PDM, whereas another may find m_2 $(m_2 \neq m_1)$ landmarks from each shape contour to construct a second PDM. Both PDMs may describe the same shape space. Even if $m_2 = m_1$, as shown in Figs. 1b and 1c, the two shape-correspondence algorithms may identify landmarks at different locations and construct two different PDMs. These two PDMs may still describe the same shape space.

TABLE 1
Compactness $C(M)$ Derived from the Shape-Correspondence Results Shown in Figs. 1a, 1b, and 1c, Respectively

M	1	2	3	4	5	6	7	8
Fig. 1a	0.5	1	1	1	1	1	N/A	N/A
Fig. 1b	0.625	1.25	1.25	1.25	1.25	1.25	1.25	1.25
Fig. 1c	0.5	1	1	1	1	1	1	1

However, the three different shape-correspondence results shown in Figs. 1a, 1b, and 1c may lead to different values in terms of the compactness, generality, and specificity measures. Tables 1, 2, and 3 indicate that the shape-correspondence result shown in Fig. 1b has a worse performance than the ones shown in Figs. 1a and 1c in terms of the compactness, generality, and specificity

TABLE 2
Generality $G(M)$ Derived from the Shape-Correspondence Results Shown in Figs. 1a, 1b, and 1c, Respectively

M	1	2	3	4	5	6	7	8
Fig. 1a	1.25	0	0	0	0	0	N/A	N/A
Fig. 1b	1.5625	0	0	0	0	0	0	0
Fig. 1c	1.25	0	0	0	0	0	0	0

TABLE 3
Specificity $S(M)$ Derived from the Shape-Correspondence Results Shown in Figs. 1a, 1b, and 1c, Respectively

M	1	2	3	4	5	6	7	8
Fig. 1a	0.10	0.31	0.31	0.31	0.31	0.31	N/A	N/A
Fig. 1b	0.13	0.39	0.39	0.39	0.39	0.39	0.39	0.39
Fig. 1c	0.10	0.31	0.31	0.31	0.31	0.31	0.31	0.31

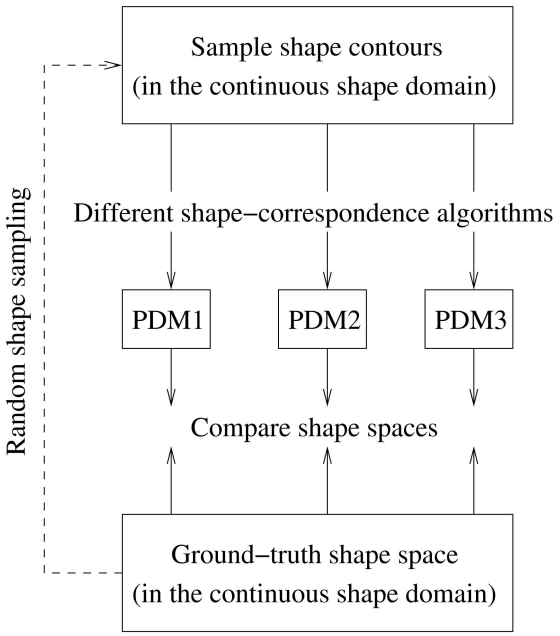


Fig. 2. Motivation for the proposed shape-correspondence benchmark.

measures. As mentioned above, all three shape-correspondence results shown in Figs. 1a, 1b, and 1c lead to the same shape space from the same set of shape contours. Therefore, these three measures may not accurately and objectively reflect the performance of a shape-correspondence result. Particularly, the shape-correspondence results shown in Figs. 1a and 1b have a different number of landmarks and the shape-correspondence results shown in Figs. 1b and 1c have the same number of landmarks. This example shows that these three measures may not reflect their relative performance in either case: They lead to the same shape space but show different values of compactness, generality, and specificity.

3 Proposed Shape-Correspondence Benchmark

In this section, we address the limitations of the compactness, generality, and specificity measures described in Section 2.3 by introducing a new benchmark to more objectively evaluate the shape-correspondence performance. In essence, a shape correspondence result is good if it leads to a PDM that can well describe the underlying shape space, which is defined in the continuous shape domain. However, the compactness, generality, and specificity measures are dependent on the number of identified landmarks and their locations.

To address this problem, our approach evaluates shape correspondence by checking whether it leads to a PDM that accurately describes the underlying shape space in the continuous shape domain. As shown in Fig. 2, our approach is based on a given ground-truth shape space. We then randomly sample a set of continuous shape contours from this shape space (as shown by the dashed line in Fig. 2). Finally, we test the shape-correspondence algorithms using these sample shape contours and evaluate their performance by assessing how well their derived PDMs describe the ground-truth shape space. To achieve our goal, two important problems must be addressed: 1) the representation and random sampling of the ground-truth shape space and 2) assessing how well a derived PDM describes the ground-truth shape space. Note that different PDMs built with different shape-correspondence methods may describe the same shape space, as shown by the example provided in Section 2.3.

In this paper, the ground-truth shape space is defined by a PDM $\mathcal{N}(\bar{\mathbf{v}}^t, \mathbf{D}^t)$, which can be arbitrarily selected (covariance matrix $\mathbf{D}^t$ must be positive definite). The shape contours and their probability density in the ground-truth shape space can be simulated by randomly sampling this PDM, generating shape vectors and then interpolating their landmarks into continuous shape contours. For simplicity, we refer to this PDM $\mathcal{N}(\bar{\mathbf{v}}^t, \mathbf{D}^t)$ as the *ground-truth PDM*.

The system diagram for the proposed benchmark is illustrated in Fig. 3. The benchmark consists of the following five components:

C1. specifying a ground-truth PDM $\mathcal{N}(\bar{\mathbf{v}}^t, \mathbf{D}^t)$ to describe the underlying ground-truth shape space,
C2. using this ground-truth PDM to randomly generate a set of continuous shape contours $S_1, S_2, \ldots, S_n$,
C3. running a test shape-correspondence algorithm on these shape contours to identify a set of corresponded landmarks,
C4. deriving a PDM $\mathcal{N}(\bar{\mathbf{v}}, \mathbf{D})$ from the identified landmark sets using (1), and
C5. assessing how well the derived PDM $\mathcal{N}(\bar{\mathbf{v}}, \mathbf{D})$ describes the ground-truth shape space defined by the ground-truth PDM $\mathcal{N}(\bar{\mathbf{v}}^t, \mathbf{D}^t)$.

This five-step process evaluates a shape-correspondence algorithm's ability to recover the underlying ground-truth shape space in the continuous shape domain.

Components C3 and C4 have been discussed in detail earlier in Section 2. For Component C1, ideally, we can specify the ground-truth PDM by arbitrarily selecting any mean shape vector $\mathbf{v}^t$ and a covariance matrix $\mathbf{D}^t$. In Section 4, we choose six ground-truth PDMs that resemble shapes of some real structures for experiments. Note that these six ground-truth PDMs may not accurately describe these real structures from an anatomic and biomedical perspective. In the following, we focus on developing algorithms to accomplish Components C2 and C5.

3.1 Generating Shape Contours for Testing Shape-Correspondence Algorithms

Given a ground-truth PDM $\mathcal{N}(\bar{\mathbf{v}}^t, \mathbf{D}^t)$ defined by k landmarks, we can randomly generate as many (landmark-based) shape vectors $\mathbf{v}_a^t$, $a = 1, 2, \ldots, n$, as necessary. More specifically, with $\mathbf{p}_j^t$ and λ_j^t, $j = 1, 2, \ldots, 2k$, being the eigenvectors and eigenvalues of $\mathbf{D}^t$, we can generate random synthetic shape vectors in the form of

$$\mathbf{v}^t = \bar{\mathbf{v}}^t + \sum_{j=1}^{2k} b_j^t \mathbf{p}_j^t, \tag{2}$$

where b_j^t is independently and randomly sampled from the 1D Gaussian distribution $\mathcal{N}(0, \lambda_j^t)$, $j = 1, 2, \ldots, 2k$.

From each generated shape vector $\mathbf{v}_a^t$, $a = 1, 2, \ldots, n$, we can derive k landmarks $\{\mathbf{v}_{a1}^t, \mathbf{v}_{a2}^t, \ldots, \mathbf{v}_{ak}^t\}$. By assuming that these k landmarks are always sequentially sampled from a continuous shape contour, we can estimate this continuous shape contour S_a' by landmark interpolation. For constructing a closed shape contour, the interpolation fills the portion

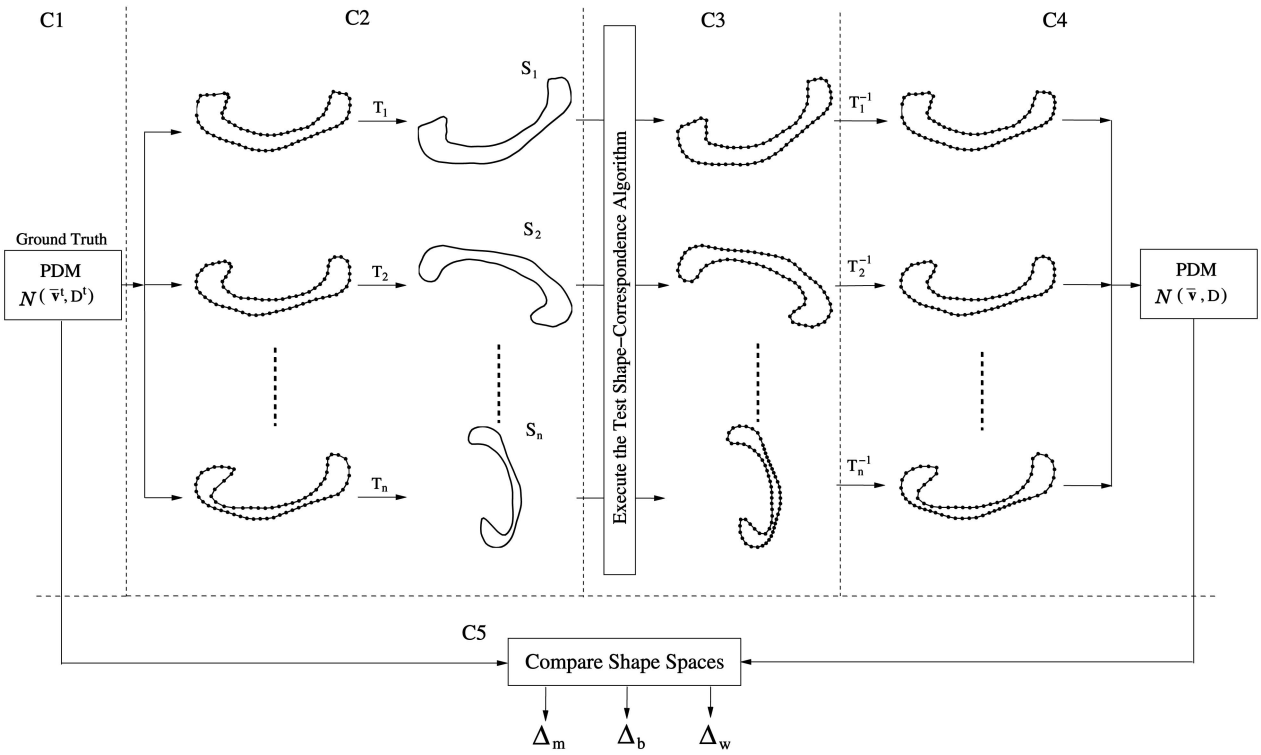


Fig. 3. An illustration of the proposed shape-correspondence evaluation benchmark.

between the last landmark $\mathbf{v}_{ak}^t$ and the first landmark $\mathbf{v}_{a1}^t$. For constructing an open shape contour, the interpolation does not fill the portion between $\mathbf{v}_{ak}^t$ and $\mathbf{v}_{a1}^t$. Although we can use any interpolation technique, we use the Catmull-Rom cubic spline [4] for all of our experiments. If the landmarks are sufficiently dense to represent the underlying shape contour (this is usually required for shape correspondence [19]), we expect that different interpolation techniques do not introduce significant differences to the proposed shape-correspondence evaluation.[2]

However, it is still not suitable to directly output S'_a, $a = 1, 2, \ldots, n$, as the sample shape contours for testing a shape-correspondence algorithm. The reasons are twofold. First, contour S'_a is a Catmull-Rom spline that is described by control points $\{\mathbf{v}_{a1}^t, \mathbf{v}_{a2}^t, \ldots, \mathbf{v}_{ak}^t\}$. We cannot directly pass these control points to the test shape-correspondence algorithm since they are corresponded landmark sets for the n shape contours. Otherwise, the test shape-correspondence algorithm can cheat by simply taking these control points as the identified corresponded landmark sets. Second, there are no rotation, scaling, and translation transformations among these n shape contours. In practice, such affine transformations are very common and their removal should be handled by the shape-correspondence algorithm.

To address these problems, we first resample each shape contour S'_a to construct a new set of control points which still represent S'_a accurately. This resampling is conducted independently on each S'_a and the newly constructed control points are not corresponded from one shape contour to another. In fact, the number of new control points from one shape contour may even be different from the number of new control points from another shape contour. In this paper, we densely resample each shape contour S'_a uniformly to construct the new control points. Second, for the newly sampled control points on each shape contour S'_a, we apply a different random affine transformation T_a. These affine transformations consist of a random rotation, a random uniform scaling, and a random translation. We then use the Catmull-Rom spline to interpolate these transformed control points and define the resulting continuous contour to be the shape contour S_a. We record the affine transformations T_a, $a = 1, 2, \ldots, n$, and then pass $S_1, S_2, \ldots, S_n$ (in fact, their control points) to the test shape-correspondence algorithm.

Note that the recorded affine transformations T_a, $a = 1, 2, \ldots, n$, are not passed to the test shape-correspondence algorithm (Component C3). If the test shape-correspondence algorithm introduces further transformations, such as Procrustes analysis, on these shape contours $S_1, S_2, \ldots, S_n$ during shape correspondence, we record and undo these transformations before outputting the shape-correspondence result. This ensures that the corresponded landmarks identified by the test shape-correspondence algorithm are placed directly back onto the input shape contours $S_1, S_2, \ldots, S_n$. Then, in Component C4, we directly apply the inverse transform T_a^{-1}, $a = 1, 2, \ldots, n$, to the landmarks identified on S_a. This guarantees the correct removal of the random affine transformation T_a before PDM construction in Component C4.

3.2 New Measures for Shape-Correspondence Performance

From the shape-correspondence results, we can derive a new PDM $\mathcal{N}(\bar{\mathbf{v}}, \mathbf{D})$ using (1). In this section, we develop three new measures for evaluating shape-correspondence

2. In Section 3.2, we define shape-correspondence evaluation measures using the Jaccard and the Dice coefficients, which are not sensitive to specific landmark-interpolation techniques.

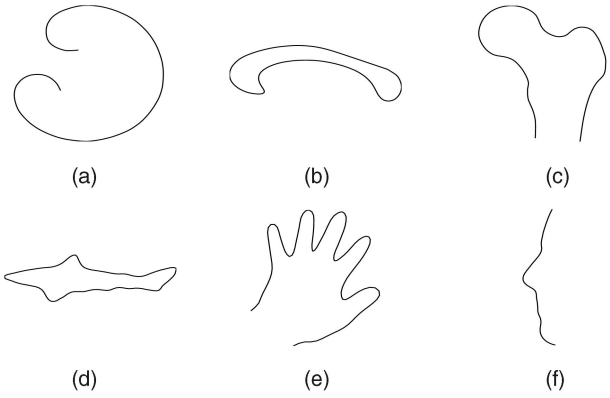


Fig. 4. The continuous shape contours $\bar{S}^t$ interpolated from the mean shape vectors $\bar{\mathbf{v}}^t$ for the six ground-truth PDMs used in our experiments. They resemble the shape of (a) kidney, (b) corpus callosum, (c) femur, (d) shark, (e) hand, and (f) face silhouette. Note that they may not accurately describe these real structures from an anatomic and biomedical perspective.

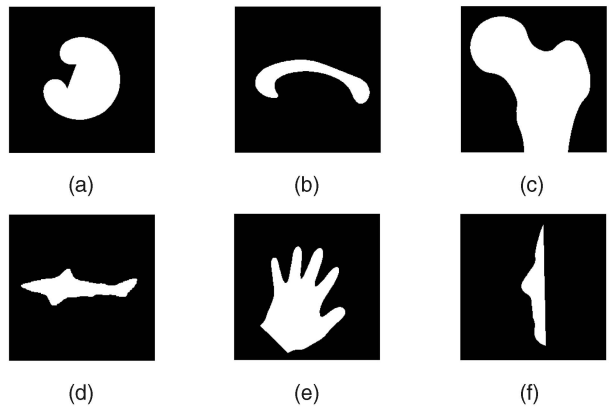


Fig. 5. Enclosed regions $R(\bar{S}^t)$ (shown in white) derived from the shape contours shown in Fig. 4. Among them, (a), (c), (e), and (f) are the enclosed regions derived from open contours by connecting the two endpoints.

performance by assessing how well the derived PDM $\mathcal{N}(\bar{\mathbf{v}}, \mathbf{D})$ describes the ground-truth shape space, which is defined by the ground-truth PDM $\mathcal{N}(\bar{\mathbf{v}}^t, \mathbf{D}^t)$. Note that we cannot directly compare the two Gaussian distributions $\mathcal{N}(\bar{\mathbf{v}}, \mathbf{D})$ and $\mathcal{N}(\bar{\mathbf{v}}^t, \mathbf{D}^t)$ using a standard technique, such as the Kullback-Leibler (K-L) distance. This is because the vectors $\bar{\mathbf{v}}$ and $\bar{\mathbf{v}}^t$ may not be comparable.[3] Instead, we assess the similarity of the shape spaces defined by these two PDMs.

We develop three measures to achieve our goal. First, we interpolate the mean shape vectors $\bar{\mathbf{v}}$ and $\bar{\mathbf{v}}^t$ into continuous mean shape contours $\bar{S}$ and $\bar{S}^t$ using the Catmull-Rom cubic spline. We then measure the difference between two continuous shape contours using the widely known Jaccard coefficient:

$$\Delta_m \triangleq \Delta(\bar{S}, \bar{S}^t) = 1 - \frac{|R(\bar{S}) \cap R(\bar{S}^t)|}{|R(\bar{S}) \cup R(\bar{S}^t)|}, \tag{3}$$

where $R(S)$ is the region enclosed by the contour S and $|R|$ computes the area of the region R. Fig. 4 shows the shape contours interpolated from the mean shape vectors $\bar{\mathbf{v}}^t$ for the six ground-truth PDMs used in our experiments and Fig. 5 illustrates their enclosed regions $R(\bar{S}^t)$. We can see that this mean-shape difference is limited to a value between [0, 1], with 0 indicating that $\bar{S}$ is exactly the same as $\bar{S}^t$. A smaller value of Δ_m indicates a better shape-correspondence result.

However, $\Delta_m = 0$ does not ensure that the shape spaces underlying these two PDMs are also the same. We use a random simulation method to further assess the differences between these two shape spaces. Specifically, we randomly generate a large set of N continuous shape contours from each PDM using (2), followed by landmark interpolation. We denote the N continuous shape contours generated from PDM $\mathcal{N}(\bar{\mathbf{v}}, \mathbf{D})$ to be $S_1^c, S_2^c, \ldots, S_N^c$[4] and the N continuous shape contours from the ground-truth PDM $\mathcal{N}(\bar{\mathbf{v}}^t, \mathbf{D}^t)$ to be $S_1^t, S_2^t, \ldots, S_N^t$. Given these two sets of continuous shape contours, $\{S_i^c\}_{i=1}^N$ and $\{S_j^t\}_{j=1}^N$, we can measure the difference between any pair of shape contours

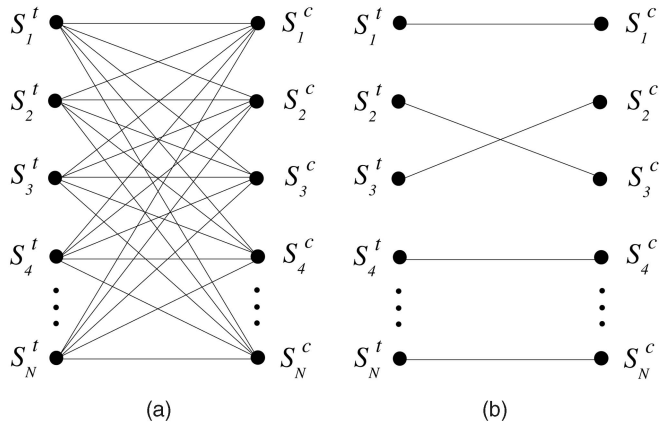


Fig. 6. An illustration of the bipartite-matching algorithm. (a) A fully connected bipartite graph where the vertices on the left side represent $\{S_j^t\}_{j=1}^N$ and the vertices on the right represent $\{S_i^c\}_{i=1}^N$. The edge weight between S_i^c and S_j^t is defined as the Jaccard-coefficient difference $\Delta(S_i^c, S_j^t)$. (b) An example result of bipartite matching.

$\Delta(S_i^c, S_j^t)$, where $i, j = 1, 2, \ldots, N$, using (3). When N is sufficiently large, the difference between these two sets of continuous shape contours can well reflect the difference in the shape spaces underlying these two PDMs. The problem is then how to measure the difference between these two sets of continuous shape contours using the Jaccard-coefficient difference measure between any pair of individual shape contours.

We define the second measure using the bipartite-matching difference between $\{S_i^c\}_{i=1}^N$ and $\{S_j^t\}_{j=1}^N$. This finds a one-to-one matching by minimizing the total Jaccard-coefficient difference between these two sets of shape contours. The calculation of this measure consists of the following steps:

1. Construct a fully connected bipartite graph, as shown in Fig. 6a, that contains $2N$ vertices for the shape contours $\{S_i^c\}_{i=1}^N$ and $\{S_j^t\}_{j=1}^N$.
2. The edge weight between each pair of vertices is defined as the Jaccard difference based on (3) between the two corresponding shape contours.
3. Apply the bipartite-matching algorithm to locate the one-to-one matching with a minimum total edge

3. The dimensions of $\bar{\mathbf{v}}$ and $\bar{\mathbf{v}}^t$ may be different, such as the PDMs derived from the landmarks shown in Figs. 1a and 1b.

4. S_i^c is used instead of S_a, $a = 1, 2, \ldots, n$, to represent the generated shape contours from PDM $\mathcal{N}(\bar{\mathbf{v}}, \mathbf{D})$ because S_a, $a = 1, 2, \ldots, n$, has been previously used in Section 3.1.

weights, as shown in Fig. 6b. We then define the bipartite-matching difference measure as

$$\Delta_b \triangleq \frac{\sum_{i=1}^{N} \Delta(S_i^c, S_{b(i)}^t)}{N},$$

where S_i^c and $S_{b(i)}^t$ are the matched pair of shape contours found by the bipartite-matching algorithm.

The bipartite-matching difference measure contains a normalization over N. This ensures Δ_b to be a value in [0, 1]. $\Delta_b = 0$ implies that the two shape spaces defined by the ground-truth PDM $\mathcal{N}(\bar{\mathbf{v}}^t, \mathbf{D}^t)$ and the derived PDM $\mathcal{N}(\bar{\mathbf{v}}, \mathbf{D})$ are identical. $\Delta_b = 1$ implies that these two PDMs describe two completely different shape spaces. Using the bipartite-matching algorithm, the measure Δ_b assesses not only whether the two shape spaces contain similar shape contours but also whether a shape contour has the same or similar probability density in these two shape spaces.

We define the third measure based on the Wald-Wolfowitz test [11] which is originally used to decide whether two sets of 1D numbers are generated from the same distribution. The basic idea is to sort all of these numbers into a sorted list and count the pairs of neighboring numbers that come from the same set. According to the Wald-Wolfowitz test, the fewer such kinds of pairs there are, the more likely it is that these two sets of 1D numbers are generated by the same distribution. However, we need to assess whether two sets of continuous shape contours $\{S_i^c\}_{i=1}^N$ and $\{S_j^t\}_{j=1}^N$ are from the same shape space. Unlike 1D numbers, the shape contours cannot be directly sorted. To address this shape-contour sorting problem, we use the generalized Wald-Wolfowitz test [11] based on the minimum spanning tree (MST) algorithm:

1. Construct a fully connected undirected graph defined by $2N$ vertices that represent the $2N$ shape contours $\{S_i^c\}_{i=1}^N$ and $\{S_j^t\}_{j=1}^N$.
2. An edge weight is defined as the Jaccard-coefficient difference based on (3) between the two shape contours connected by this edge. Note that the edges are constructed not only across the set $\{S_i^c\}_{i=1}^N$ and the set $\{S_j^t\}_{j=1}^N$ but also within each of these sets. Therefore, we need to calculate not only $\Delta(S_i^c, S_j^t)$ but also $\Delta(S_i^c, S_j^c)$ and $\Delta(S_i^t, S_j^t)$, $i, j = 1, 2, \ldots, N$.
3. Find the MST of the constructed graph [7]. An MST of a graph is defined to be the spanning tree with the minimum total edge weight and a spanning tree of a graph is a tree subgraph that contains all of the vertices.
4. From the $2N - 1$ edges in the resulting MST, count the total number of edges that connect two vertices that represent two shape contours within either $\{S_i^c\}_{i=1}^N$ or $\{S_j^t\}_{j=1}^N$, respectively. We denote this number as W. Note that the maximum possible value for W is $2N - 2$. We can then define the Wald-Wolfowitz difference measure as

$$\Delta_w \triangleq \frac{W}{2N-2}.$$

According to the generalized Wald-Wolfowitz test, the smaller the Δ_w value is, the more likely it is that the $2N$ shape contours $\{S_i^c\}_{i=1}^N$ and $\{S_j^t\}_{j=1}^N$ are generated from the same shape space.

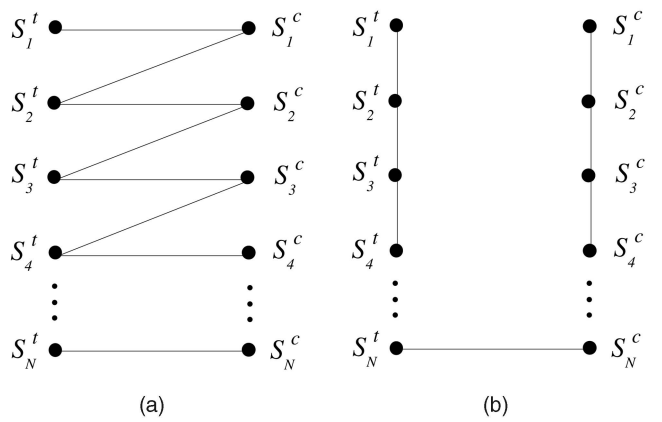


Fig. 7. An illustration of two possible MSTs found from the graph constructed from the $2N$ shape contours $\{S_i^c\}_{i=1}^N$ and $\{S_j^t\}_{j=1}^N$. (a) An MST with $\Delta_w = 0$ and (b) an MST with $\Delta_w = 1$.

Like the bipartite-matching difference measure, the Wald-Wolfowitz difference measure contains a normalization by $2N - 2$. This ensures Δ_w to be a value in [0, 1]. Fig. 7a shows an example of an MST with $\Delta_w = 0$ and Fig. 7b shows an example of an MST with $\Delta_w = 1$. We can see that, when $\Delta_w = 0$, the two sets of shape contours $\{S_i^c\}_{i=1}^N$ and $\{S_j^t\}_{j=1}^N$ cannot be well separated by the MST-based sorting. This indicates that they are more likely to be from the same shape space. Like Δ_b, Δ_w assesses not only whether the two shape spaces contain similar shape contours but also whether a shape contour has the same or similar probability density in these two shape spaces.

Note that both Δ_b and Δ_w are calculated using $2N$ randomly generated shape contours from PDMs $\mathcal{N}(\bar{\mathbf{v}}, \mathbf{D})$ and $\mathcal{N}(\bar{\mathbf{v}}^t, \mathbf{D}^t)$. Different rounds of random generations will most likely result in different $2N$ shape contours and therefore may lead to different values of Δ_b and Δ_w. To ensure the statistical reliability of these two difference measures, we repeat the random generation of the $2N$ shape contours for many rounds and take the average Δ_b and Δ_w to evaluate shape-correspondence performance.

These three difference measures Δ_m, Δ_b, and Δ_w are based on the Jaccard coefficient for measuring the differences between two shape contours. It is well known that the Jaccard coefficient is not the only difference measure between two contours. Another widely used difference measure is the Dice coefficient, defined as

$$\Delta'(\bar{S}, \bar{S}^t) = 1 - \frac{2 \cdot |R(\bar{S}) \cap R(\bar{S}^t)|}{|R(\bar{S})| + |R(\bar{S}^t)|}, \tag{4}$$

which also takes value in [0, 1]. It is straightforward to use this Dice coefficient instead of the Jaccard coefficient when calculating Δ_m, Δ_b, and Δ_w. In Section 4, we provide experiments using both the Jaccard and the Dice coefficients.

4 Experiments

In this section, the proposed benchmark is used to evaluate five 2D shape-correspondence algorithms:

Richardson and Wang's implementation of an algorithm that combines landmark sliding, insertion, and deletion (SDI) [19], Thodberg's implementation of the minimum description length algorithm (T-MDL) [22], [21], Karlsson and Ericsson's implementation of the MDL algorithm (E-MDL) [14], Karlsson and Ericsson'ss implementation of the MDL algorithm with curvature distance minimization (E-MDL+CUR) [14], and Karlsson and Ericsson's implementation of the reparameterization algorithm by minimizing euclidean distance (EUC) [14].

As discussed in Section 3, the ground-truth PDM can be arbitrarily specified. For our experiments, we construct the ground-truth PDM that "resembles" certain types of real structures. Specifically, we collect real shape contours of kidney, corpus callosum (callosum for short), femur, shark, hand, and human face silhouette and then apply any shape-correspondence algorithm on them.[5] The shape correspondence results are then normalized using Procrustes analysis [13] and, finally, six ground-truth PDMs are constructed using (1). In constructing these six ground-truth PDMs, we identify 128 corresponded landmarks along each shape contour. Therefore, their mean shape vectors have a dimension of 256 and their covariance matrices have a dimension of 256×256. Each of these six ground-truth PDMs defines a ground-truth shape space, which is the start point for our evaluation, as indicated in Component C1 in Fig. 3. Note that these ground-truth PDMs may not accurately describe the shape space underlying these six real structures from an anatomic or biomedical perspective and the proposed benchmark tests the shape-correspondence algorithms on synthetic shape contours. Because of this, we use the term "resemble" when referring to these six ground-truth PDMs. In Section 5, we further discuss a possible modification of the proposed benchmark so that ground truth can more accurately describe the shape of real structures.

In Component C2, we randomly generate $n = 800$ synthetic shape contours S_a, $a = 1, 2, \ldots, n$, from each of the six ground-truth PDMs that are passed to the test shape-correspondence algorithm. As discussed in Section 3.1, this includes a dense uniform resampling followed by a random affine transformation T_a consisting of a random rotation, a random translation, and a random uniform scaling. Specifically, we uniformly resample each shape contour, generating approximately 25,600 landmarks as the new control points. We randomly rotate each shape contour about its center of mass, where the rotation angle is uniformly distributed in [0, 360] degrees. We randomly translate each shape contour by moving its center of mass (x_c, y_c) to $(\rho x_c, \rho y_c)$, where ρ is a scalar uniformly distributed in [0, 1]. We randomly scale each shape contour with a scaling factor uniformly distributed in [0.75, 1], where the scaling factor along the x and y directions is the same.

In Component C3, we set the expected number of corresponded landmarks to be 64 for the test shape-correspondence algorithms. The reasons are twofold: 1) T-MDL requires the number of corresponded landmarks identified from each shape contour to be a power of two and 2) we intentionally set the number of landmarks identified by the test shape-correspondence algorithm to be different from the ground-truth PDM. Note that this choice precludes the use of landmark-based measures, such as the K-L distance, for comparing the two PDMs. For E-MDL, E-MDL+CUR, T-MDL, and EUC, we additionally set the maximum number of iterations to be 20. For E-MDL, E-MDL+CUR, and EUC, we use the recommended settings. For T-MDL, we use the recommended settings for both closed and open shape contours; however, we do not allow the endpoints to move for the open shape contours. Note that both E-MDL and T-MDL are based on a simplified description length cost function and are considered to be only approximate versions of MDL [9]. Although E-MDL and T-MDL use the same simplified MDL cost function, they are implemented in different ways. The most notable difference is that E-MDL does not distinguish between open and closed shapes, whereas T-MDL does [22], [10].

When determining the values of Δ_b and Δ_w in Component C5, we randomly generate $N = 2,000$ shape contours. This random simulation process is repeated for 50 rounds to check the stability of Δ_b and Δ_w. We use Goldberg and Kennedy's cost scaling push relabeling algorithm [12] to calculate the bipartite matching for Δ_b and Kruskal's algorithm [7] to calculate the MST for Δ_w.

Figs. 8 and 9 and Table 4 show the evaluation results based on the Jaccard coefficient (3). Both of the values of Δ_b and Δ_w shown in Table 4 are the average values over the 50 rounds of random simulation. In Figs. 8 and 9, we can see that the values of Δ_b and Δ_w are relatively stable over the 50 rounds of random simulation. Figs. 8 and 9 and Table 4 additionally show that, when comparing the performance of two shape-correspondence algorithms on a given ground-truth PDM, the evaluation results in terms of Δ_m, Δ_b, and Δ_w are consistent if the values of Δ_m resulting from these two methods are quite different. For example, when applying SDI and T-MDL on the ground-truth PDM that resembles the kidney, T-MDL has worse performance in terms of Δ_m, Δ_b, and Δ_w. However, if two shape-correspondence algorithms produce similar values of Δ_m when applied to a given ground-truth PDM, the evaluation results in terms of Δ_m, Δ_b, and Δ_w may be inconsistent. For example, when applying T-MDL and E-MDL on the ground-truth PDM that resembles the femur, T-MDL performs slightly better than E-MDL in terms of Δ_m, but E-MDL performs better than T-MDL in terms of Δ_b and Δ_w. In this case, we believe that the evaluation results in terms of Δ_b and Δ_w are more accurate and reliable because they assess the similarity of the shape spaces, whereas Δ_m only assesses the similarity of the mean shape contours. Therefore, we focus on using Δ_b and Δ_w to evaluate the shape-correspondence performance.

From Figs. 8 and 9 and Table 4, we can see that E-MDL demonstrates the best performance for the ground-truth PDMs that resemble the callosum, shark, hand, and face silhouette. EUC and E-MDL demonstrate the best performance for the ground-truth PDMs that resemble the kidney and femur. SDI performs less favorably than E-MDL, E-MDL+CUR, EUC, and T-MDL for the ground-truth PDMs that resemble the femur and shark, whereas T-MDL performs less favorably than E-MDL, E-MDL+CUR, EUC, and SDI for the ground-truth PDMs that resemble the kidney, callosum, and face silhouette. SDI and EUC perform less favorably than E-MDL, T-MDL and E-MDL+CUR

5. In our experiments, this is achieved by manually labeling one corresponded landmark on each contour and then performing a uniform sampling for remaining landmarks along the shape contour. For open shape contours, such as kidney and femur, we assume the endpoints are corresponded across all of the shape contours and, therefore, manual labeling is not required.

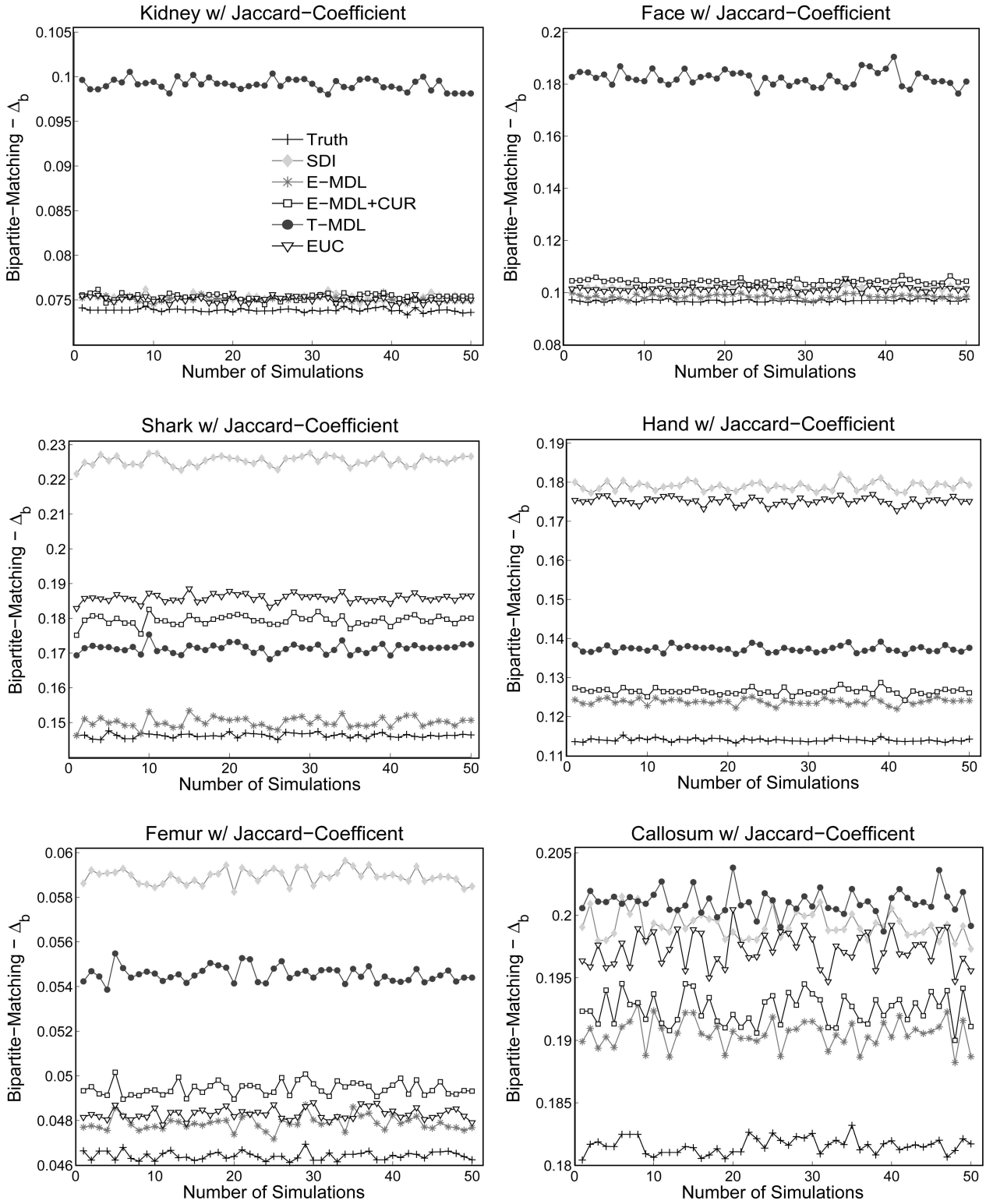


Fig. 8. The values of Δ_b for the six ground-truth PDMs resulting from the five test shape-correspondence algorithms using the Jaccard coefficient. The *x*-axis indicates the round of the random simulation. The curves with the "+" symbols are the values of Δ_b between each ground-truth PDM and itself.

for the ground-truth PDM that resembles the hand. In general, E-MDL demonstrates the best overall performance for the six ground-truth PDMs in terms of Δ_b and Δ_w and this is largely consistent with the evaluation results reported in [14].

One problem is to justify the scale of these three measures. Δ_m has a straightforward geometric explanation: the region coincidence between two closed contours. When $\Delta_m = 0$, the (interpolated continuous) mean shape contours of the two PDMs are exactly the same. However, for Δ_b and Δ_w shown in

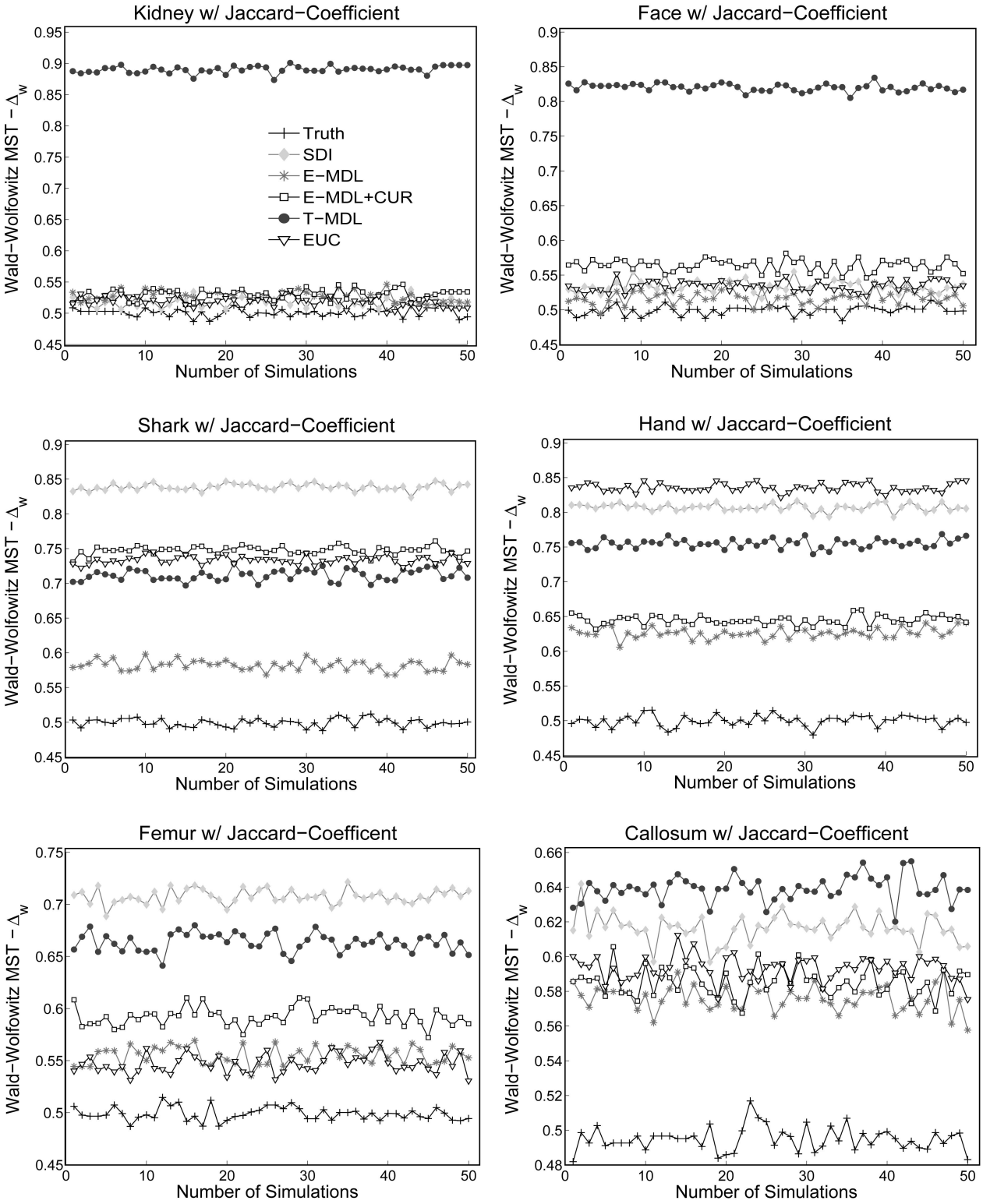


Fig. 9. The values of Δ_w for the six ground-truth PDMs resulting from the five test shape-correspondence algorithms using the Jaccard coefficient. The *x*-axis indicates the round of the random simulation. The curves with the "+" symbols are the values of Δ_w between each ground-truth PDM and itself.

Table 4, there is no such direct geometric explanation. For example, in Table 4, the average Δ_b value for the ground-truth PDM that resembles the kidney is 0.09913 when using T-MDL. It is not clear whether this number indicates a big or small difference between the shape spaces defined by these two PDMs. In the following, we propose two strategies to quantify the scales of Δ_b and Δ_w.

First, we calculate Δ_b and Δ_w between a ground-truth PDM $\mathcal{N}(\bar{\mathbf{v}}^t, \mathbf{D}^t)$ and itself. Clearly, the randomness in the simulation makes both Δ_b and Δ_w to be nonzero, as shown by

TABLE 4
The Values of Δ_b and Δ_w Resulting from the Five Test Shape-Correspondence Algorithms Using the Jaccard Coefficient

	Kidney			Callosum			Femur		
Algorithm	Δ_m	Δ_b	Δ_w	Δ_m	Δ_b	Δ_w	Δ_m	Δ_b	Δ_w
SDI	0.01096	0.07528	0.51891	0.05131	0.19916	0.61644	0.01147	0.05893	0.70761
T-MDL	0.03454	0.09913	0.89017	0.08339	0.20105	0.63913	0.00831	0.05455	0.66374
E-MDL	0.01317	0.07508	0.52706	0.08104	0.19053	0.57594	0.00844	0.04788	0.55502
E-MDL+CUR	0.01249	0.07533	0.53036	0.06887	0.19248	0.58517	0.01131	0.04943	0.59226
EUC	0.01084	0.07508	0.51833	0.04750	0.19727	0.59323	0.00558	0.04830	0.54688
	Shark			Hand			Face		
Algorithm	Δ_m	Δ_b	Δ_w	Δ_m	Δ_b	Δ_w	Δ_m	Δ_b	Δ_w
SDI	0.09409	0.22534	0.83869	0.04060	0.17897	0.80723	0.00629	0.10189	0.53390
T-MDL	0.03669	0.17134	0.71172	0.06479	0.13733	0.75525	0.14074	0.18226	0.82021
E-MDL	0.02377	0.15011	0.58209	0.04795	0.12381	0.62636	0.01485	0.09847	0.51671
E-MDL+CUR	0.07434	0.17942	0.74742	0.05216	0.12646	0.64483	0.02724	0.10432	0.56390
EUC	0.04706	0.18588	0.73365	0.09425	0.17520	0.83579	0.00815	0.10148	0.53419

TABLE 5
Several Settings for Measuring the Scales of the Δ_b and Δ_w Values

		Kidney	Callosum	Femur	Shark	Hand	Face
Setting 0	Δ_b	0.07386	0.18158	0.04642	0.14627	0.11400	0.09691
	Δ_w	0.50136	0.49493	0.49889	0.49939	0.50005	0.49923
Setting 1	Δ_m	0.00553	0.00943	0.00069	0.01456	0.00026	0.00421
	Δ_b	0.07402	0.18147	0.04641	0.14674	0.11396	0.09705
	Δ_w	0.50541	0.49634	0.49862	0.51060	0.49854	0.50075
Setting 2	Δ_m	0.05319	0.06860	0.00713	0.12935	0.00277	0.05759
	Δ_b	0.08604	0.19014	0.04682	0.19899	0.11403	0.10742
	Δ_w	0.85595	0.58888	0.51585	0.97910	0.50077	0.56030
Setting 3	Δ_b	0.07840	0.19337	0.05001	0.15772	0.12224	0.10509
	Δ_w	0.51565	0.51031	0.51563	0.50765	0.51043	0.50855
Setting 4	Δ_b	0.09156	0.22982	0.06051	0.19471	0.14737	0.13392
	Δ_w	0.58022	0.57667	0.58608	0.56868	0.57069	0.56204

the curves with "+" symbols in Figs. 8 and 9. The average Δ_b and Δ_w values (over the 50 rounds of random simulations) are shown in Table 5 under the row "Setting 0." The values of Δ_b and Δ_w between a ground-truth PDM and itself provides a lower bound in evaluating a shape-correspondence algorithm for this ground-truth PDM. In Figs. 8 and 9, we can see that the values of Δ_b and Δ_w resulting from all shape-correspondence algorithms are larger than this lower bound.

Second, we slightly modify a ground-truth PDM and then calculate the values of Δ_b and Δ_w between the modified ground-truth PDM and the original ground-truth PDM. Specifically, we apply two modifications to a ground-truth PDM: 1) Adding a constant value α to all of the landmarks in $\bar{\mathbf{v}}^t$. This is in fact a translation of the mean shape vector, 2) multiplying all $\sqrt{\lambda_j^t}$, $j = 1, 2, \ldots, 2k$, by a constant factor β, where λ_j^t, $j = 1, 2, \ldots, 2k$, are the eigenvalues of $\mathbf{D}^t$. This uniformly scales the standard deviation along each of the principal directions. Settings 1, 2, 3, and 4 in Table 5 show the resulting values of Δ_b and Δ_w by choosing different α and β. In Setting 1, we choose $\alpha = 0.1\% \cdot r(\bar{\mathbf{v}}^t)$ for all six ground-truth PDMs, where $r(\bar{\mathbf{v}}^t)$ is the *mean-shape radius* of the respective ground-truth PDM that is defined as the average euclidean distance from each landmark in $\bar{\mathbf{v}}^t$ to its center of mass. In Setting 2, we choose $\alpha = 1.0\% \cdot r(\bar{\mathbf{v}}^t)$ for all six ground-truth PDMs. In Settings 3 and 4, we choose $\beta = 1.1$ and $\beta = 1.3$, respectively, for all six ground-truth PDMs. We choose $\beta = 1$ for Settings 1 and 2 and $\alpha = 0$ for Settings 3 and 4.

Table 5 provides some general concepts about the scale of the Δ_b and Δ_w values shown in Table 4. For example, $\Delta_b = 0.19727$ when applying EUC on the ground-truth PDM that resembles the callosum, as shown in Table 4. This value of Δ_b is similar to the value of $\Delta_b = 0.19014$ resulting from Setting 2 in Table 5. We also find that the values of Δ_b and Δ_w are more sensitive to the translation of the mean shape vector (the values of α are very small in Settings 1 and 2) than to the scaling along the principal directions. This indicates that, when the resulting Δ_m is very large for a test shape-correspondence algorithm, we can safely say that its performance is poor.

Figs. 10 and 11 and Table 6 show the evaluation results based on the Dice coefficient (4). Like the values shown in Table 4, the values of Δ_b and Δ_w shown in Table 6 are the

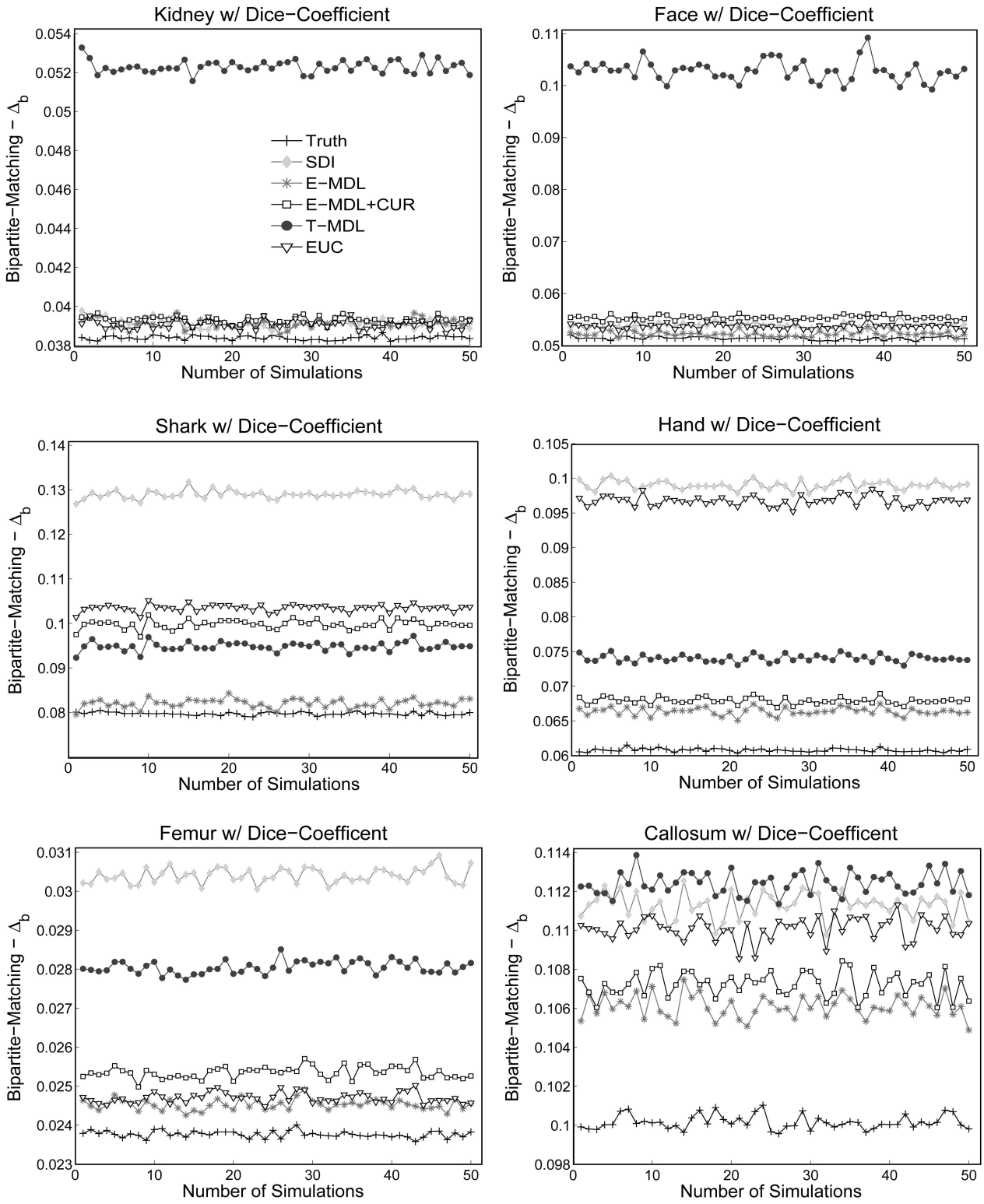


Fig. 10. The values of Δ_b for the six ground-truth PDMs resulting from the five test shape-correspondence algorithms using the Dice coefficient. The *x*-axis indicates the round of the random simulation. The curves with the "+" symbols are the values of Δ_b between each ground-truth PDM and itself.

average values over the 50 rounds of random simulations. Comparing the Δ_b results shown in Fig. 8 to the Δ_b results shown in Fig. 10 and the Δ_w results shown in Fig. 9 to the Δ_w results shown in Fig. 11, the values using the Dice coefficient are largely consistent with those using the Jaccard coefficient. For example, among all five test algorithms, E-MDL has the best overall performance in terms of the Δ_m, Δ_b, and Δ_w using either the Jaccard or the Dice coefficients. As a result, the average Δ_w values shown in Tables 4 and Table 6 are similar since the Wald-Wolfowitz test only counts the number of MST edges,

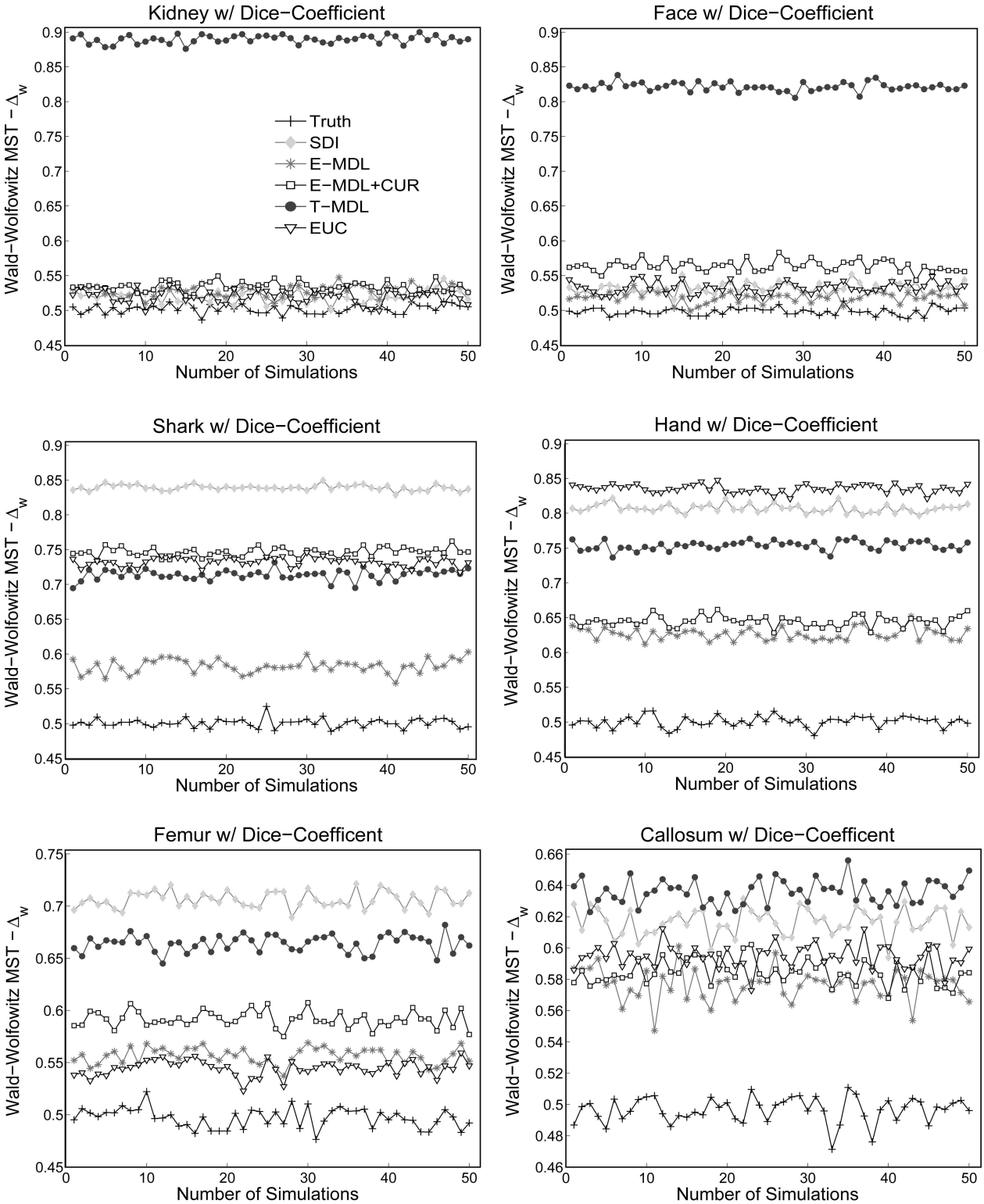


Fig. 11. The values of Δ_w for the six ground-truth PDMs resulting from the five test shape-correspondence algorithms using the Dice coefficient. The *x*-axis indicates the round of the random simulation. The curves with the "+" symbols are the values of Δ_w between each ground-truth PDM and itself.

whereas the bipartite-matching test depends on the specific values of the Jaccard and the Dice coefficients.

Finally, runtime is another important issue in evaluating a shape-correspondence algorithm. Table 7 shows the CPU time (in seconds) taken by the five shape-correspondence algorithms to correspond the $n = 800$ random shape contours generated by each ground-truth PDM. These experiments are run on a Linux PC with a 3.4 GHz 2M L2 Xeon processor with 4 Gbytes of RAM. Note that all five test shape-correspondence algorithms are implemented using Matlab. We can

TABLE 6
The Values of Δ_b and Δ_w Resulting from the Five Test Shape-Correspondence Algorithms Using the Dice Coefficient

	Kidney			Callosum			Femur		
Algorithm	Δ_m	Δ_b	Δ_w	Δ_m	Δ_b	Δ_w	Δ_m	Δ_b	Δ_w
SDI	0.00551	0.03916	0.52005	0.02605	0.11127	0.61653	0.00584	0.03040	0.70624
T-MDL	0.01757	0.05230	0.88958	0.04351	0.11243	0.63600	0.00412	0.02805	0.66427
E-MDL	0.00663	0.03914	0.52805	0.04213	0.10608	0.57646	0.00424	0.02452	0.55659
E-MDL+CUR	0.00628	0.03930	0.53388	0.03557	0.10723	0.58441	0.00569	0.02535	0.59153
EUC	0.00545	0.03908	0.51996	0.02451	0.11009	0.59488	0.00280	0.02470	0.54453
	Shark			Hand			Face		
Algorithm	Δ_m	Δ_b	Δ_w	Δ_m	Δ_b	Δ_w	Δ_m	Δ_b	Δ_w
SDI	0.04937	0.12894	0.83911	0.02086	0.09904	0.80711	0.00292	0.05386	0.53203
T-MDL	0.01869	0.09482	0.71387	0.03345	0.07403	0.75363	0.07570	0.10283	0.82146
E-MDL	0.01203	0.08208	0.58272	0.02440	0.06637	0.62633	0.00748	0.05224	0.51871
E-MDL+CUR	0.03861	0.09976	0.74652	0.02665	0.06791	0.64506	0.01357	0.05538	0.56377
EUC	0.02410	0.10348	0.73266	0.04947	0.09676	0.83556	0.00374	0.05374	0.53293

see that SDI takes the least CPU time to accomplish shape correspondence on all six ground-truth PDMs. However, in terms of Δ_m, Δ_b, and Δ_w, the shape-correspondence results produced by SDI are not as favorable as E-MDL. This may indicate that the shape-correspondence problem has a high complexity and may require a longer runtime to achieve better accuracy.

5 A Note on Benchmarking with Real Structures

As mentioned above, the ground-truth PDM can be arbitrarily specified in the proposed benchmark. In Section 4, we select six PDMs that resemble several real structures, including kidney, callosum, femur, shark, hand, and face silhouette. These six ground-truth PDMs may not exactly describe the real shape of the kidney, callosum, femur, shark, hand, and face silhouette from an anatomic or biomedical perspective and the shape contours used for testing shape correspondence are synthetically generated.

In Fig. 12, we show a possible modification of the proposed benchmark to evaluate shape correspondence by assessing its ability to construct a PDM that accurately describes the shape space of a real structure. This is accomplished by introducing a new way to define the ground-truth shape space in Component C1: We do not use a ground-truth PDM since the shape space of a real structure may not be well described by a Gaussian-distribution-based PDM. Instead, we use a large collection of real shape contours $\{S_1^r, S_2^r, \ldots, S_N^r\}$ to represent the shape space of the real structure. We also assume that there is no scaling, translation, and rotation transformation among these N real shape contours. In Component C2, we select a subset of n shape contours from the N shape contours for testing shape correspondence. The Δ_m measure does not apply since we do not know the mean shape contour in the ground-truth shape space. The Δ_b and Δ_w measures can still be used since they compare two sets of the shape contours. In this modification, one set of contours is $\{S_1^r, S_2^r, \ldots, S_N^r\}$, and the other is $\{S_1^c, S_2^c, \ldots, S_N^c\}$ generated from the PDM constructed in Component (C4).

In practice, obtaining a large set of real shape contours is difficult and labor intensive, especially if we want to extend this modified benchmark to 3D. Note that this may not be impossible because we only need to collect the real shape contours $\{S_1^r, S_2^r, \ldots, S_N^r\}$ and do not need to label the corresponded landmarks as ground truth. Another difficulty is that, in this modification, the collected set of real shape contours $\{S_1^r, S_2^r, \ldots, S_N^r\}$ must represent the shape space of the real structure. The proposed benchmark described in Section 3 avoids both difficulties by introducing a ground-truth PDM that resembles but may not accurately describe, a real structural shape. From the ground-truth PDM, we can generate as many synthetic shape contours as we want.

TABLE 7
CPU Time (in Seconds) that Are Taken by the Five Shape-Correspondence Algorithms

Algorithm	SDI	T-MDL	E-MDL	E-MDL+CUR	EUC
Kidney	734	52140	106942	260938	28534
Callosum	703	44732	107506	278832	28420
Femur	740	59663	109875	261093	28538
Shark	745	44937	106277	270933	27115
Hand	739	50784	107317	304504	29572
Face	745	50710	103822	259551	28286

6 Conclusion

In conclusion, this paper introduced a new benchmark to evaluate the performance of landmark-based shape correspondence algorithms and their derived PDMs. Differently from previous shape-correspondence evaluation methods, we start from a given (ground truth) PDM, which specifies a ground-truth shape space. In this benchmark, we randomly sampled a set of shape contours from this ground-truth shape space for testing a shape-correspondence algorithm. We then evaluated the performance of the test shape-correspondence algorithm by checking whether it led to a PDM that describes the ground-truth shape space well. Three new measures were introduced to quantify the difference between two shape spaces. We also applied the

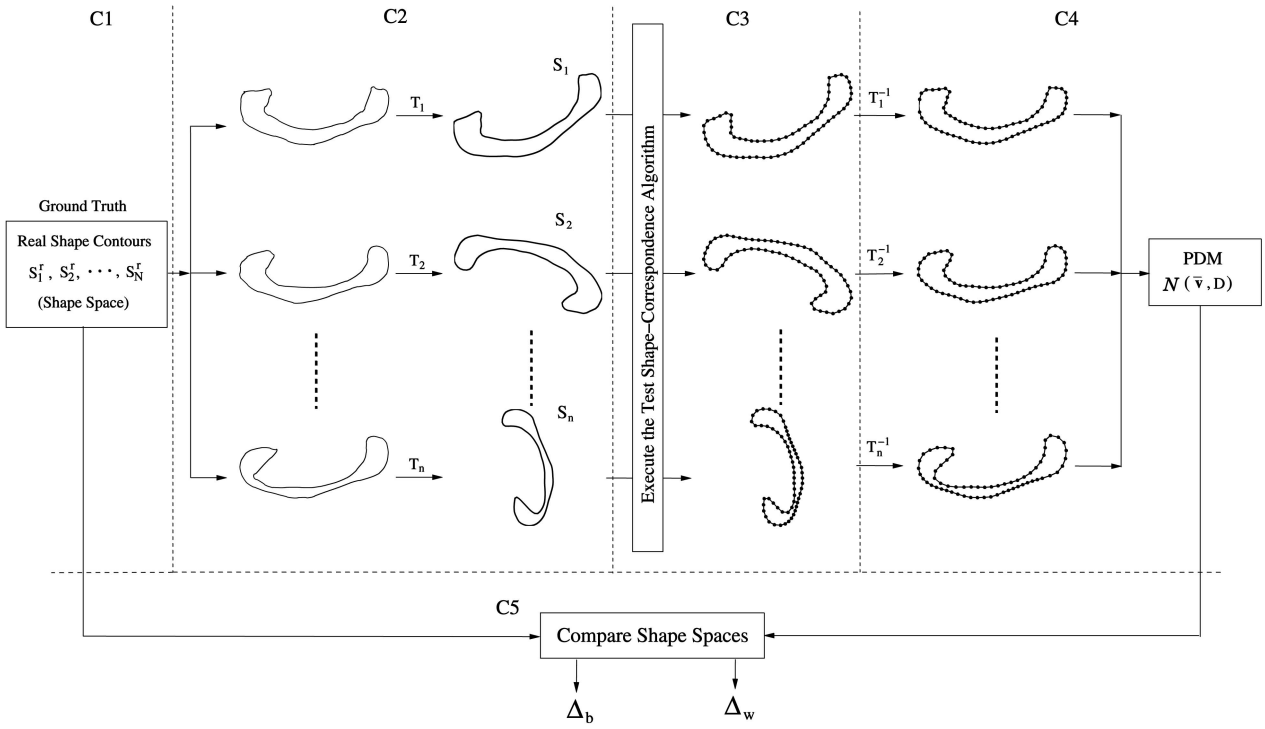


Fig. 12. An illustration of a possible modification to the proposed benchmark to evaluate shape correspondence by assessing its ability to construct a PDM that accurately describes the shape space of a real structure.

benchmark to evaluate five available 2D shape correspondence algorithms. By introducing a ground-truth PDM, we believe the proposed benchmark allows for a more objective evaluation of shape correspondence performance that is landmark independent. The proposed benchmark can easily be extended to 3D cases by computing the shape-contour difference (3) using the intersection and union of the volumes bounded by the two shape surfaces.

ACKNOWLEDGMENTS

This work was funded, in part, by NSF-EIA-0312861 and AFOSR FA9550-07-1-0250. The authors thank T. Thodberg for providing the package of T-MDL, A. Ericsson for providing the packages of E-MDL, E-MDL+CUR, and EUC, and T. Richardson for providing the package of SDI. They thank A. Goldberg and R. Kennedy for providing the bipartite-matching software package. They would like to thank Eva Czabarka and Jijun Tang for their invaluable suggestions. Some preliminary results of this paper were published in a conference proceeding [17].

REFERENCES

[1] S. Belongie, J. Malik, and J. Puzicha, "Shape Matching and Object Recognition Using Shape Contexts," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 4, pp. 509-522, Apr. 2002.

[2] F. Bookstein, "Principal Warps: Thin-Plate Splines and the Decomposition of Deformations," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 11, no. 6, pp. 567-585, June 1989.

[3] F. Bookstein, "Landmark Methods for Forms without Landmarks: Morphometrics of Group Differences in Outline Shape," *Medical Image Analysis,* vol. 1, no. 3, pp. 225-243, 1997.

[4] E. Catmull and R. Rom, "A Class of Local Interpolating Splines," *Computer Aided Geometric Design,* pp. 317-326, 1974.

[5] H. Chui and A. Rangarajan, "A New Point Matching Algorithm for Non-Rigid Registration," *Computer Vision and Image Understanding,* vol. 89, pp. 114-141, 2003.

[6] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active Shape Models—Their Training and Application," *Computer Vision and Image Understanding,* vol. 61, no. 1, pp. 38-59, Jan. 1995.

[7] T. Cormen, C. Leiserson, and R. Rivest, *Introduction to Algorithms.* MIT Press, 1990.

[8] R. Davies, "Learning Shape: Optimal Models for Analysing Natural Variablility," PhD dissertation, Univ. of Manchester, 2002.

[9] R. Davies, C. Twining, T. Cootes, J. Waterton, and C. Taylor, "A Minimum Description Length Approach to Statistical Shape Modeling," *IEEE Trans. Medical Imaging,* vol. 21, no. 5, pp. 525-537, May 2002.

[10] A. Ericsson, "Automatic Shape Modelling with Applications in Medical Imaging," PhD dissertation, Centre for Math. Sciences, Lund Univ., 2006.

[11] J. Friedman and L. Rafsky, "Multivariate Generalization of the Wald-Wolfowitz and Smirnov Two-Sample Tests," *Annals of Statistics,* vol. 7, pp. 697-717, 1979.

[12] A. Goldberg and R. Kennedy, "An Efficient Cost Scaling Algorithm for the Assignment Problem," *Math. Programming,* vol. 71, pp. 153-178, 1995.

[13] J. Gower and G. Dijksterhuis, *Procrustes Problems.* Oxford Univ. Press, 2004.

[14] J. Karlsson and A. Ericsson, "A Ground Truth Correspondence Measure for Benchmarking," *Proc. Int'l Conf. Pattern Recognition,* vol. 3, pp. 568-573, 2006.

[15] A. Kotcheff and C.J. Taylor, "Automatic Construction of Eigen-Shape Models by Genetic Algorithm," *Proc. Information Processing in Medical Imaging Conf.,* pp. 1-14, 1997.

[16] M. Leventon, E. Grimson, and O. Faugeras, "Statistical Shape Influence in Geodesic Active Contours," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 316-323, 2000.

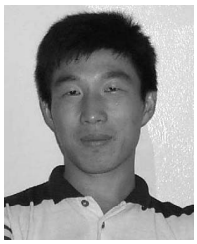
[17] B. Munsell, P. Dalal, and S. Wang, "A New Benchmark for Shape Correspondence Evaluation," *Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention,* vol. 1, pp. 507-514, 2007.

[18] M. Powell, "A Thin Plate Spline Method for Mapping Curves into Curves in Two Dimensions," *Proc. Computational Techniques and Applications,* pp. 43-57, 1995.

[19] T. Richardson and S. Wang, "Nonrigid Shape Correspondence Using Landmark Sliding, Insertion and Deletion," *Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention,* vol. 2, pp. 435-442, 2005.
[20] M. Styner, K. Rajamani, L.-P. Nolte, G. Zsemlye, G. Szekely, C. Taylor, and R. Davies, "Evaluation of 3D Correspondence Methods for Model Building," *Proc. Information Processing in Medical Imaging Conf.,* pp. 63-75, 2003.
[21] H. Thodberg, "Adding Curvature to Minimum Description Length Shape Models," *Proc. British Machine Vision Conf.,* vol. 2, pp. 251-260, 2003.
[22] H. Thodberg, "Minimum Description Length Shape and Appearance Models," *Proc. Information Processing in Medical Imaging Conf.,* pp. 51-62, 2003.
[23] S. Wang, T. Kubota, and T. Richardson, "Shape Correspondence through Landmark Sliding," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 143-150, 2004.
[24] J. Xie and P. Heng, "Shape Modeling Using Automatic Landmarking," *Proc. Int'l Conf. Medical Image Computing and Computer Assisted Intervention,* vol. 2, pp. 709-716, 2005.

Brent C. Munsell received the BS degree in electrical engineering from Michigan State University in 1994 and the ME degree in electrical engineering from Clemson University in 2004. From 1994 to 1999, he was a lieutenant in the US Navy stationed on the USS Enterprise and, from 1999 to 2006, he worked at Scientific Research Corp. as an engineer. He is currently a PhD candidate in the Department of Computer Science and Engineering at the University of South Carolina. His research interests include statistical shape analysis and medical image processing. He is a student member of the IEEE.

Pahal Dalal received the BE degree in computer engineering from the University of Mumbai in 2005 and the ME degree in computer science and engineering from the University of South Carolina in 2007. He is currently a PhD candidate in the Department of Computer Science and Engineering at the University of South Carolina. His research interests include computer vision and medical imaging, including the areas of statistical shape analysis and medical image segmentation. He is a student member of the IEEE.



Song Wang received the PhD degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC) in 2002. From 1998 to 2002, he also worked as a research assistant in the Image Formation and Processing Group at the Beckman Institute at UIUC. In 2002, he joined the Department of Computer Science and Engineering at the University of South Carolina, where he is currently an associate professor. His research interests include computer vision, medical image processing, and machine learning. He is a member of the IEEE and the IEEE Computer Society.