Shape Comparison Using Perturbing Shape Registration

Yifeng Jiang[1], Erin Edmiston[2], Fei Wang[2], Hilary P. Blumberg[1,2]
Lawrence H. Staib[1,3,4], Xenophon Papademetris[1,3]
Department of [1]Diagnostic Radiology, [2]Psychiatry, [3]Biomedical Engineering, [4]Electrical Engineering
Yale University, New Haven, CT, USA
yifeng.jiang@yale.edu

Abstract

Shape registration *is often involved in computing statistical differences between groups of shapes, which is a key aspect of morphometric study. The results of shape difference are found to be sensitive to registration,* i.e.*, different registration methods lead to varied results. This raises the question of how to improve the reliability of registration procedures. This paper proposes a perturbation scheme, which perturbs registrations by feeding them with different resampled shape groups, and then aggregates the resulting shape differences. Experiments are conducted using three typical registration algorithms on both synthetic and biomedical shapes, where more reliable inter-group shape differences are found under the proposed scheme.*

1. Introduction

Evidence suggests that morphological difference in anatomical structures often reflect the underlying functional variation. With the development of medical imaging and image analysis techniques, morphometric study is becoming one of the most accessible approaches for investigating the functions of biological structure, quantifying development, and identifying pathology, potentially leading to better diagnosis and treatment.

A key scenario in morphometric analysis is to compute the morphological difference between two groups of shapes. In the current literature, there are two ways to quantify this shape difference. One is by statistical significance testing [2], where the resulting p-value or z-score represents the difference. The other is by estimating shape classifiers [12], whose classification ratios are used as the measurement. In this case, the shape difference can also be visualized by deforming the mean shape along the orthogonal direction of a classifier [13]. However, in many real applications it is difficult to find a practical useful classifier, and as a result, significance testing has been adopted widely [21].

For either significance testing or classifier estimation, discrete descriptors need to be obtained for the original contours or surfaces. Historically, many location-invariant measurements, such as lengths, angles, areas and volumes, have been used [3]. Descriptors that reflect a shape's global properties, like Fourier descriptors, spherical harmonic functions, and wavelet descriptors, have also been employed. They can represent the full information of the original shape, but most of them are not able to compare localized shape differences [1] (Some recent works using spherical harmonics and wavelet [6, 20] can do this but registration is also required). Thus, descriptors such as dense landmarks and deformation fields are preferred in many cases. However, to obtain such descriptors, shapes are required to be registered, which unfortunately still largely remains a research topic [13]. The overall procedure for computing the inter-group shape difference is briefly illustrated in Fig.1.

In this paper, we follow the simplest definition of *shape difference*: using a set of dense landmarks as shape descriptors, and for each set of corresponding landmarks, applying Hotelling's T squared testing on their coordinates to find a p-value as the local shape difference. Here shape registration is posed as first to find corresponding landmarks along different shapes, and second to align the shapes so as to eliminate the scaling, translation, and rotation. Both parts can be done by different methods. Alignment is a mature process and different methods lead to similar results [19], and in this paper we use Generalized Procrustes Analysis [9]; However, for the first part–registration, different methods result in significantly dissimilar correspondences and consequentially to varied shape differences. This may not be surprising because as an ill-posed problem, its solution really depends on the prior/regulation that assumed in each method. Given the shape data and a collection of registration methods, which result is the most trustworthy? This is a serious question, but to our best knowledge, it has seldom been discussed before.

In order to answer this question, we need to design a cri-

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

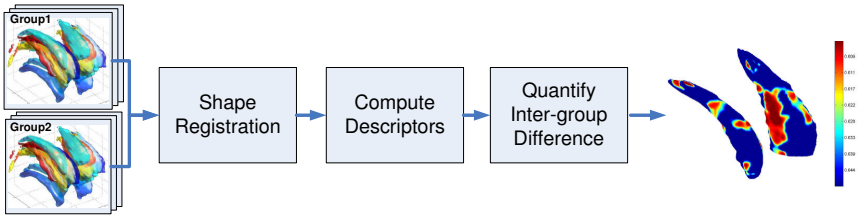

Figure 1: Three basic modules involved in computing the inter-group shape difference.

terion to compare shape registration methods. Established criteria do exist in shape *generative analysis* [7], such as model generalization error, model specificity, and model compactness, *etc*. However, it is hard to tell whether they can also be applied for *discriminative analysis*. Generative analysis concerns the determination of a compact rule to represent different shapes. The landmarks are positioned so that shapes look as similar as possible to each other. This process may ignore subtle differences among the shapes in spite of their potential value in discriminative analysis. In shape discriminative analysis, the *classification ratio* might be the most sounding candidate as a criterion. However, our extensive experiments found that this is doubtful. Numerous classifiers and feature selection methods exist, and evaluations based on different combinations often do not conform to each other. Furthermore, in some cases, worse registration leads to better classification ratios for many classifiers. The pseudo shape differences introduced by incorrect registration could serve as good discriminative features and improve the classification ratio.

So here we turn to a closely related but less difficult question: Given shape data and a registration method, how can we find more reliable results of shape comparison? This paper tries to answer it by developing a universal scheme to reduce the sensitivity of the computation of shape differences to the registration procedures. The intuition is that shape registrations are often unstable, affected by either their parameters or the data under registration. Such procedures have a good chance to be improved by perturbing themselves and aggregating the results.

The remainder of this paper is organized as follows: Section 2 introduces the perturbation scheme, goes over a conceptual justification for it, and describes a detailed implementation; In Section 3 we conduct some preliminary experiments and demonstrate the results on both synthetic and real biomedical data. Section 4 concludes the paper with a brief discussion.

2. Method

2.1. A perturbation scheme for registration

We considered two ways to perturb a registration procedure. One is to perturb the parameters of registration, which include the initial conditions. The other is to perturb the shape data – most shape registration algorithms are data-driven processes. If two shapes are inputted into an algorithm with different other shapes, the resultant correspondence between them normally changes.

This paper chooses the latter one because: (1) Parameters are very diverse among different algorithms. It is difficult to design a unified scheme to perturb parameters for different algorithms. However, the scheme of resampling the input data can be identical. (2) Given a group of shapes, a certain registration algorithm performs the best with a certain set of parameters. Deviating parameters from their optima could result in less accurate or incorrect correspondence. It is unlikely for a sub-optimal or even wrong correspondence to make a positive contribution in aggregation. On the other hand, if the distribution of the resampled data is similar to the original one, it is reasonable to believe that the "true" shape differences are well preserved, so each registration is pursuing the same objective whose results have a good chance to be integrated to find a more reliable estimation. The schemes based on perturbing the input data has also been successfully adopted in machine learning [4,8,11] to improve the performance of classifiers.

When the ground truth of shape difference is given, it is very straightforward to evaluate the computed shape difference. In this paper, a group of synthetic bump boxes are generated for this purpose. For real biomedical data whose true difference is never known, we compare the results from different registration methods: the more similar they are, the more reliable the computations are considered – This idea has also been adopted in image registration literature [10, 15], to evaluate the registration accuracy when no ground truth exists.

2.2. Why perturbation works

Given a group of shapes $\mathcal{S} = \{\mathbf{s}_n, n = 1, \ldots, N\}$, the goal of shape registration is to find the correspondence across all $\{\mathbf{s}_n\}$. One explicit way to represent the correspondence is locating a set of corresponding landmarks $\mathbf{x}_n = \{\boldsymbol{x}_{n,m} = (x_{n,m}, y_{n,m}), m = 1, \ldots, M\}$ (in 2D cases) for each shape $\mathbf{s}_n$. For simplicity, we assume there exists a mean shape $\mathbf{s}_0$ (not necessarily in the given $\mathcal{S}$) whose landmarks are given and fixed as $\mathbf{x}_0$. In this case, the

registration procedure is for each $\mathbf{s}_n$ to find M landmarks $\mathbf{x}_n$ which correspond to $\mathbf{x}_0$.

Registration is a data-driven process. If we denote the correspondence resulting from a certain registration procedure as ϕ, we can write $\mathbf{x}_n = \phi(\mathbf{s}_n, \mathcal{S})$. This is true when a group-wise registration algorithm is adopted, where the resulting correspondence is effected by every member of $\mathcal{S}$. In other words, any change, say, by excluding, or including any single $\mathbf{s}$, would make the whole group of $\{\mathbf{x}_n\}$ different. It is also true when a pair-wise registration algorithm is adopted, in which case we need to select a template shape and register all the others to this template. The registration results are often largely affected by the template, while the template shape is selected from all the members in $\mathcal{S}$.

To evaluate the effect of perturbation, we consider a sequence of shape groups $\{\mathcal{S}_k\}$ each consisting of N independent observations from the same underlying distributions as $\mathcal{S}$. The simplest way to get the final correspondence from perturbation is to replace $\phi(\mathbf{s}_n, \mathcal{S})$ by the average of $\phi(\mathbf{s}_n, \mathcal{S}_k)$ over k, which ideally is $\phi_A(\mathbf{s}) = E_S\phi(\mathbf{s}, \mathcal{S})$, where E_S denotes the expectation over $\mathcal{S}$, and the subscript A in ϕ_A denotes aggregation. Here we do a simple comparison between $\phi(\mathbf{s}, \mathcal{S})$ and $\phi_A(\mathbf{s})$, which is similar to the Bagging procedure in classifier aggregation [4].

Assume the joint distribution of $(\mathbf{x}, \mathbf{s})$ to be P once $\mathbf{x}_0$ is given. Let $(\mathbf{x}, \mathbf{s})$ be independently drawn from P, and take $\mathbf{s}$ to be a fixed input value and $\mathbf{x}$ an output value. Then

$$E_S(\mathbf{x} - \phi(\mathbf{s}, \mathcal{S}))^2 = \mathbf{x}^2 - 2\mathbf{x}E_S\phi(\mathbf{s}, \mathcal{S}) + E_S\phi^2(\mathbf{s}, \mathcal{S}). \quad (1)$$

Using $E_S\phi(\mathbf{s}, \mathcal{S}) = \phi_A(\mathbf{s})$ and applying the inequality $EX^2 \geq (EX)^2$ to the third term in (1) gives

$$E_S(\mathbf{x} - \phi(\mathbf{s}, \mathcal{S}))^2 \geq (\mathbf{x} - \phi_A(\mathbf{s}))^2. \quad (2)$$

Integrating both sides of (2) over the joint distribution of $(\mathbf{x}, \mathbf{s})$, we get that the mean-squared error of $\phi_A(\mathbf{s})$ is lower than the mean-square error averaged over $\mathcal{S}$ of $\phi(\mathbf{s}, \mathcal{S})$, which means statistically $\phi_A(\mathbf{s})$ will yield better correspondence than $\phi(\mathbf{s}, \mathcal{S})$.

How much better depends on the degree of inequality:

$$E_S\phi^2(\mathbf{s}, \mathcal{S}) \geq [E_S\phi(\mathbf{s}, \mathcal{S})]^2 . \quad (3)$$

This is apparently affected by the instability of $\phi(\mathbf{s}, \mathcal{S})$. If $\phi(\mathbf{s}, \mathcal{S})$ does not change too much with replicate $\mathcal{S}$ the two sides will be nearly equal, and aggregation will not help. The more highly variable the $\phi(\mathbf{s}, \mathcal{S})$ are, the more improvement aggregation may produce.

In practice, we only have a single shape group $\mathcal{S}$ without the lavish replicates $\{\mathcal{S}_k\}$. To imitate the perturbation procedure, we can take repeated bootstrap samples, $\{\mathcal{S}^{(B)}\}$, and take ϕ_B as $\phi_B = E_B\phi(\mathbf{s}, \mathcal{S}^{(B)})$. In this case, we are considering $\phi_B(\mathbf{s}) = \phi_A(\mathbf{s}, P_S)$ instead of $\phi_A(\mathbf{s}, P)$, where P_S is the distribution that concentrates mass $1/N$ at each point $(\mathbf{x}_n, \mathbf{s}_n)$. Then ϕ_B is caught in two currents: On the one hand, if the procedure is unstable, it can give improvement through aggregation; On the other hand, if the procedure is stable, then $\phi_B = \phi_A(\mathbf{s}, P_S)$ will not be as accurate for data drawn from P as $\phi_A(\mathbf{s}, P)$. There is a cross-over point between instability and stability at which ϕ_B stops improving on $\phi(\mathbf{s}, \mathcal{S})$ and does worse.

Once the correspondence is established, the inter-group statistical shape difference we discuss in this paper is also determined since significance testing is a fixed routine. As mentioned in Section1, a simple computation of shape difference is to find a p-value for each set of corresponding landmarks $\{\boldsymbol{x}_{n,m}, \forall n\}$ by doing Hotelling T^2 Testing on their coordinates. In other words, $\{p_m, m = 1 \ldots M\}$ are fully determined by $\{\boldsymbol{x}_{n,m}\}$, which can be written as $\{p_m\} = \mathbf{P}(\{\boldsymbol{x}_{n,m}\}) = \mathbf{P}(\{\phi(\mathbf{s}_n, \mathcal{S}_k)\})$. By applying the same inequality as above, we may also have

$$E_S\left[\mathbf{P}(\{\mathbf{x}_n\}) - \mathbf{P}(\{\phi(\mathbf{s}_n, \mathcal{S}_k)\})\right]^2 \geq \left[\mathbf{P}(\{\mathbf{x}_n\}) - (E_S(\mathbf{P}(\{\phi(\mathbf{s}_n, \mathcal{S}_k)\}))\right]^2 . \quad (4)$$

This suggests that both aggregating the correspondences before doing significance testing, and directly aggregating the testing results, have chance to improve the computation of final inter-group shape difference, which is the ultimate goal of our task. Because aggregating the correspondences[1] is more complicated than aggregating the significance testing results, the latter is adopted in this paper.

2.3. Implementation

To implement the perturbation scheme, we need to take care of a number of issues: (1) How to select the mean shape $\mathbf{s}_0$ and its $\mathbf{x}_0$? (2) How to form the resampled shape group $\{\mathcal{S}_k^{(B)}\}$? (3) How to compare shape differences? We would also like to test aggregation methods other than averaging.

2.3.1 Selecting the mean shape $\mathbf{s}_0$

To find a perfect mean $\mathbf{s}_0$ for a group of shapes is beyond the scope of this paper. Here, a "typical" shape $\mathbf{s}_0'$ is chosen to approximate $\mathbf{s}_0$, and we simply take all its points to be $\mathbf{x}_0$. This typical shape is chosen to have the least distance to all the others. The distance between two shapes can be set as the norm of the coordinate difference of a set of corresponding landmarks after Procrustes alignment. We can register the whole shape group to obtain the landmarks, but as a rough approximation, simply distributing landmarks by evenly sampling each shape contours often works well.

[1] In practice, aggregating the correspondences is not as trivial as averaging the coordinates of landmarks. The justification in this section is only a conceptual one.

Since we need to establish the correspondence between any $\mathbf{s}_n$ and $\mathbf{s}_0$, $\mathbf{s}_0'$ will join all the resampled groups $\{\mathcal{S}_k^{(B)}\}$ during registration. Furthermore, because the resulting landmarks on $\mathbf{s}_0'$ after registering $\mathcal{S}_k^{(B)}$, say $\mathbf{x}_0^{reg}$, would never automatically locate exactly at $\mathbf{x}_0$, we need to get the corresponding landmarks on $\mathbf{s}_n$ for $\mathbf{x}_0$ by interpolation. Since we are aggregating the results of significance testing as mentioned at the end of Section 2.2, we will interpolate the p-values from $\mathbf{x}_0^{reg}$ onto $\mathbf{x}_0$, which can be achieved by linear interpolation along the contour of $\mathbf{s}_0'$.

2.3.2 Forming the resampled groups $\{\mathcal{S}_k^{(B)}\}$

In this paper each $\mathcal{S}_k^{(B)}$ is formed by randomly sampling the original group with replacement. Suppose we are computing the shape difference for two groups $\mathcal{S}^A$ and $\mathcal{S}^B$, where $\mathcal{S} = \{\mathcal{S}^A, \mathcal{S}^B\}$. Each $\mathcal{S}^{(B)}$ will include shapes from both of them, where N^A samples are randomly drawn from $\mathcal{S}^A$ and N^B samples from $\mathcal{S}^B$. This is repeated for K times to obtain $\{\mathcal{S}_k^{(B)}, k = 1, \ldots, K\}$. Often N^A and N^B are chosen to be equal to the size of $\mathcal{S}^A$ and $\mathcal{S}^B$ when bootstrap replicates are sampled, but it is found in our experiments that adopting smaller N^A and N^B often leads to better results. This is probably due to reason mentioned in the last section: On the one hand, N^A and N^B both should be as big as possible, so as to reflect the shape distribution in the original group; on the other hand, they can not be too big so $\{\mathcal{S}_k^{(B)}, k = 1, \ldots, K\}$ will not be different enough from each other to perturb the registration procedures. We need to reach a balance here. In this paper, we make them around 80% of the size of original $\mathcal{S}^A$ and $\mathcal{S}^B$.

Another way to form $\mathcal{S}^{(B)}$ is also tried in this paper: For each shape $\mathbf{s}$ in $\mathcal{S}$, choose N^A of its closest shapes from $\mathcal{S}^A$ and N^B from $\mathcal{S}^B$, which will yield $\{\mathcal{S}_k^{(B)}, k = 1, \ldots, K\}$, where N is the total sample number in original shape group. Our initial intention is that this "cluster" scheme could improve the quality of shape registration because we observed that most registration algorithms tends to do a better job for shapes less deformed from each other. However, it turns out that the results are not so good as the random sampling scheme, and the reason are probably twofold, (1) the shape distribution in the sub-set $\mathcal{S}_k^{(B)}$ is very different from the original group $\mathcal{S}$; and (2) the pre-defined $\mathbf{s}_0$, which needs to be included in every $\mathcal{S}_k^{(B)}$, becomes an outlier of most sub-sets, thus can not be well registered.

2.3.3 Aggregating perturbation results

After registering all $\{\mathcal{S}_k^{(B)}\}$, we now have K sets of p values for each landmark $\boldsymbol{x}_{0,m}$, $\{p_{m,k}, m = 1, \ldots, M, k = 1, \ldots, K\}$. To aggregate them, we can:

1. Classify $\{p_{m,k}\}$ to be significant or insignificant for a given significance level α, and the final classification can be decided by a majority voting.
2. Simply let $p_m = \frac{1}{K}\sum_{k=1}^{K} p_{m,k}$, as described in Section 2.2. We denote it as p_m^1. This might be a less sounding aggregation method since p is highly nonlinear. But in our experiments, its performance has been consistently good.
3. First cluster $\{p_{m,k}, k = 1, \ldots, K\}$ into two groups, and let p_m to be the mean of the larger one, denoted as p_m^2. This is based on the observation that occasionally the registration procedure finds incorrect correspondence for some $\{\mathcal{S}_k^{(B)}\}$.
4. Conduct statistical testing on $\{p_{m,k}, k = 1, \ldots, K\}$, where the null hypothesis is $p > \alpha$ and α is the significance level. The resulting p_m is denoted as p_m^3.

In the first case, we do not have a p value for each landmark any more, and instead a label l_m is obtained, which still enables the comparisons between shape differences.

2.3.4 Comparing the resulting shape differences

Finally, to compare $\{p_m, m = 1, \ldots, M\}$ resulting from different registration methods, we need to define a similarity metric. Since p values are always used together with a significance level, α, and whether it is significant or not is more important than the actual p, here we count the number of p_ms that are both significant or insignificant in different results, *i.e.*,

$$\begin{aligned} \mathrm{sim}(\alpha) &= \frac{1}{M}\left(nr(\{m, p_m^A > \alpha\} \cap \{m, p_m^B > \alpha\})\right. \\ &+ \left. nr(\{m, p_m^A \le \alpha\} \cap \{m, p_m^B \le \alpha\})\right), \quad (5) \end{aligned}$$

where the superscripts of p indicate the registration approach. In practice, α is typically 0.05. In this paper, we compute the similarity over $\alpha \in (0.01, 0.1)$. For the first aggregation method described in the above section, we simply count the number of common labels between shape differences.

Normally, multiple comparison correction is needed after significance testing to find an appropriate α to eliminate the over-optimal results. In this paper, it is not performed since we focus on comparing shape differences, and this comparison has been carried out over a large range of α.

The full approach is described in Algorithm 1.

2.4. Registration methods

We study three typical registrations methods [5,16,18] in this paper, which assume different priors for the shape transformation and also adopt distinctive shape distance metrics. Thodberg [18] followed the Minimum Description

Algorithm 1: The proposed perturbation scheme

input : $\mathcal{S}^A, \mathcal{S}^B, K, N^A, N^B$.
output: $\{p_m^1, p_m^2, p_m^3, m = 1, \ldots, M\}$.
begin
  Find a typical shape $\mathbf{s}_0'$.
  for $k = 1$ *to* K **do**
    Form bootstrap replicates $\mathcal{S}_k$ by randomly choosing N^A, N^B examples from $\mathcal{S}^A, \mathcal{S}^B$ with replacement;
    Enclose $\mathbf{s}_0'$ into $\mathcal{S}_k$
    Register $\mathcal{S}_k$ and get landmarks set $\{\mathbf{x}_{n,k}\}$, where the landmark on $\mathbf{s}_0'$ is $\mathbf{x}_{0,k}$;
    Do significance testing for $\{\mathbf{x}_{n,k}\}$ and get $\mathbf{p}_k$
    Interpolate $\mathbf{p}_k$ from $\mathbf{x}_{0,k}$ to $\mathbf{x}_0$
  Aggregate $\{\mathbf{p}_k, k = 1 \ldots K\}$ into final $\mathbf{p}^1$, $\mathbf{p}^2$, and $\mathbf{p}^3$.
end

Length (MDL) principle in information theory to formulate the prior on transformation. The algorithm slides landmarks along the shape contours or surfaces, codes their locations (after minimizing the Procrustes distance) using Point Distribution Models (PDM) and assumes that the true correspondence will achieve the shortest code length; Instead of using any information metric, Chui *et al*. [5] adopted the bending energy derived from Thin Plate Spline (TPS) to measure the goodness of transformation. The bending energy is alternatively minimized together with a fuzzy shape distance until convergence; Huang *et al*. [16] did not measure the shape distance directly in space, instead, they converted shapes into images by distance transform, and measured the image similarity first by mutual information (MI) and then by the sum of square difference (SSD). A dense transformation field which is modeled by B-spline free-form deformation is utilized to impose a smoothness prior on the transformation. Both [5, 18] are group-wise registration algorithms. They involve estimation for the mean shape, to which every sample shape is registered during the whole registration procedure; while [16] is a pure pair-wise registration algorithm, for which a template shape needs to be chosen before registration. The implementations of [5] and [16] are publicly accessible from the authors' websites. The implementation of [18] can be requested from the author.

3. Results

We test the perturbation scheme on both synthetic shapes and real biomedical shapes. The synthetic shapes are a set of bump boxes, for which the true correspondence between shapes is given. The real biomedical shapes are a set of femur and corpus callosum profiles.

3.1. Results on synthetic shapes

As shown in Fig 2, each bump box is a rectangle box with three semi-circles on the edges. The bump locations, $\{P1, P2, P3\}$, are different from sample to sample and uniformly distributed along the corresponding edge. The whole set is divided into two groups, where the first one has larger top but smaller right bumps. A small scale of Gaussian noise is imposed on all the shapes. Fig. 3 shows the results on bump boxes. Each column shows the results obtained by a particular registration method. Here both N^A and N^B are equal to 6, which means the size of each resampled shape group is 13. 24 resampled groups are formed. For all registration methods, the perturbation scheme has resulted in shape differences that are obviously closer to the ground truth with the exception of the majority voting and p^2 resulted by MDL-based registration [18]. This exception probably implies that the algorithm implemented by [18] is less stable, for which it is difficult to decide whether p^2 should be equal to the mean of the major cluster or reverse. The similarities of p^1 shows that [18] also gains most by the perturbation among all the three registration methods. This may also indicates its instability according to the analysis in Section 2.2.

The figure also shows that the correspondences established by shape registration all look good by visual inspection. The shape generalization results (Not shown here) are also found to be very similar to that obtained by the ground truth correspondence. However, large areas of incorrect shape difference are observed, especially for the results of [5, 18], if compared with the ground truth shape difference. This suggests that the inter-group shape difference could set a much stricter or more sensitive criterion for shape registration, compared with the criteria for generalization analysis. One reason is that the difference here is defined by a p value from significance testing, which is not decided by the magnitude of any error, but only by statistical significance.

Perturbations with different size and number of resampled group were also tested and the results are not very sensitive to them. Optimal results are found with resampled group size N^A, N^B equaling 6 or 7. Once N^A, N^B exceed 9, the results remain nearly unchanged. We also tested the perturbation scheme on synthesized boxes with other number of bumps. It is found that the advantage of the perturbation scheme is more significant with boxes with fewer bumps, or in other words, with less complicated shapes.

3.2. Results on real biomedical shapes

We computed shape differences for two sets of real biomedical shapes: one contains 32 contours of femurs extracted from supine projection X-rays, which are publicly accessible from the Division of Imaging Science and

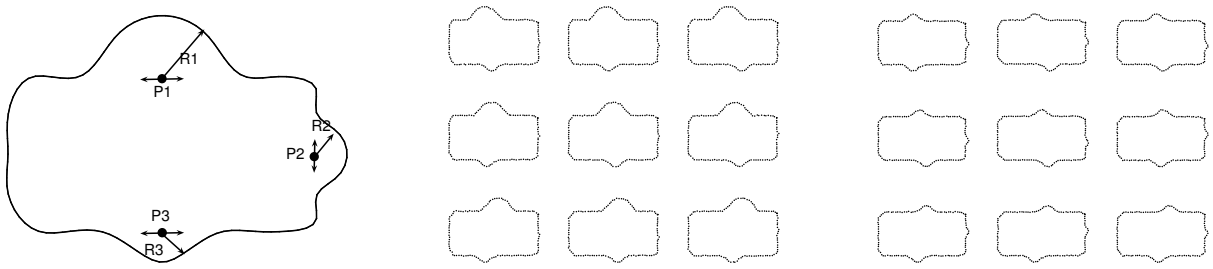

Figure 2: Two groups of three-bump boxes. $\{P1, P2, P3\}$ are uniformly distributed for both groups. $R1$ are greater in the first group, while $R2$ is greater in the second group. $R3$ is identical for both groups.

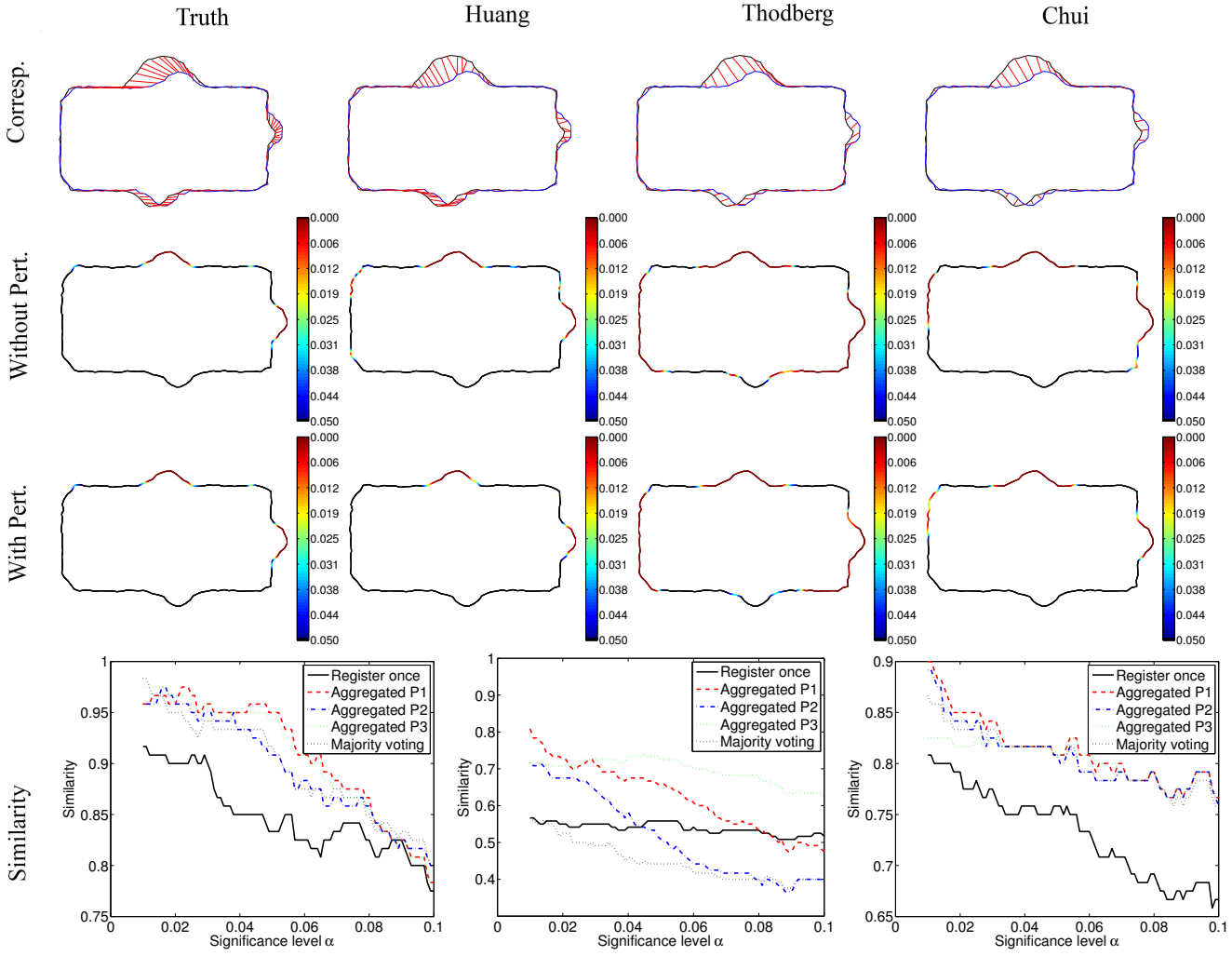

Figure 3: Results on bump boxes. Each column shows the results obtained by a particular registration. The first column is obtained by ground truth correspondence; the second, third, and forth columns are obtained by registration algorithms [16], [18] and [5], which are denoted by authors' names, 'Huang', 'Thodberg', and 'Chui' respectively. The first row shows the correspondence between two typical shape samples. The second row illustrates the shape difference computed without perturbation (register the whole shape group at once). The third row demonstrates the shape difference $\{p^1\}$ (cf. Section 2.3) computed with perturbation (In the first column, this is identical to the second row). The forth row gives the similarity between the ground truth and the shape differences computed by these three algorithms.

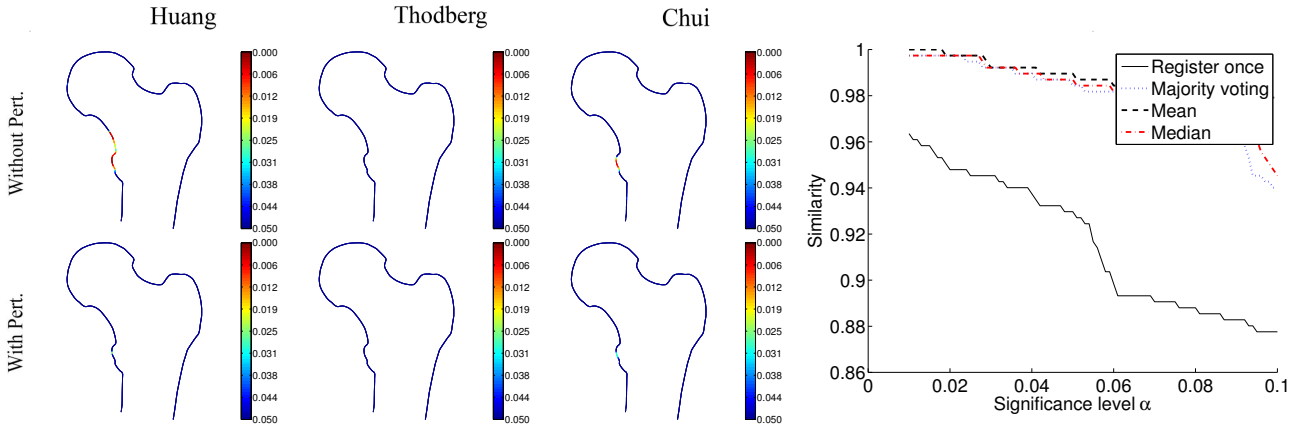

Figure 4: Shape differences of femur. From left to right, the results are computed by [16], [18], and [5] respectively. First row shows the results without perturbation while the second row shows $\{p^1\}$ obtained by perturbation. The graph at right shows the similarity between different results.

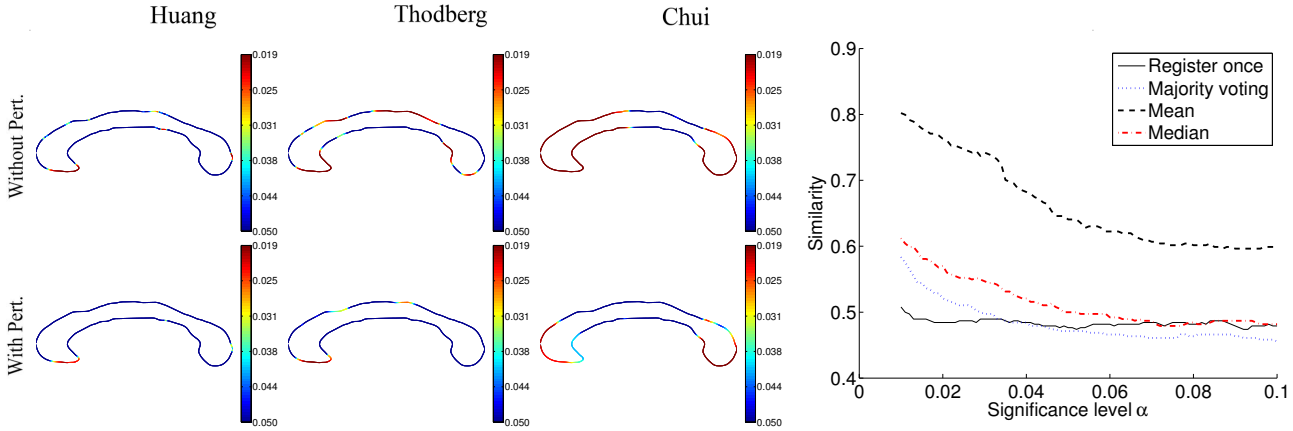

Figure 5: Shape difference of corpus callosum profiles. The arrangement of the figures are the same as in Fig. 4. Here the aggregation results from majority voting shows similar performance as those without perturbation, but results from other aggregation methods show significantly improved similarity between different registration methods.

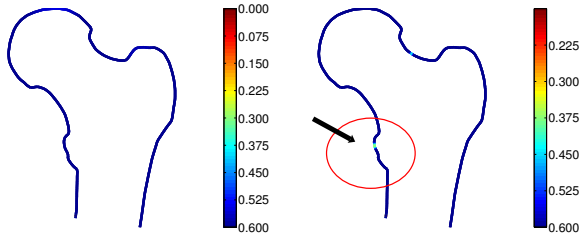

Figure 6: The femur shape difference computed by [18] when the significance level α is scaled up to 0.6 to allow less significant difference to appear. The aggregated results $\{p^2\}$ show difference on the top of lesser trochanter, while no significant result is observed without perturbation. Such a high α does not make much sense in statistics, but since we focus on comparing shape differences, this observation still provides useful hints.

Biomedical Engineering, The University of Manchester, and the other encloses 30 profiles of corpus callosum manually traced from brain MRI images. The femur shapes are divided into two groups according to the size of the lesser trochanter – a small bump just below the femur head. The first group has 6 shapes with smaller lesser trochanter. The corpus callosum shapes are also divided into two groups, each with 15 samples. They have different size anterior genus, the left most area of the corpus callosum.

We compute the sum of all the similarities between any two methods chosen from the three. As shown in Fig. 4 and 5, large differences are observed between the results from different registration methods. But this difference is very much reduced after the perturbation scheme is employed. This indicates that the perturbation scheme leads to more reliable computation of shape difference.

The perturbation method does not appear to sacrifice sensitivity to obtain more reliable shape differences. In Fig. 6, the aggregated results show more significant differences compared with registration at once.

4. Discussion and future work

This paper proposes a perturbation scheme to improve the reliability of the computation for inter-group shape difference. It provides a easy way to improve an existing registration method, since all that is needed is to add a loop in front to feed different resampled groups to the registration procedures, and a back end that aggregates all the results. Both experimental and theoretical evidence have shown that the method can gain increased reliability in computation.

The proposed scheme works for both pair-wise and group-wise methods. For pair-wise registration algorithms, it is also found that simply perturbing the registration template would achieve a similar effect (only tested for [16]). This scheme would not work for the basic SPHARM-based registration methods [17], where each shape is registered to a circle/sphere individually. However, it should be applicable to some more recent methods [6, 14], where registration between different shapes is involved.

Although only 2D shapes are examined in this paper, the proposed scheme is a generic one and can be directly applied for 3D shapes, which will be studied in the future. We would also like to examine the effects of perturbation on registration and significance testing separately, which are largely considered together in this paper. The effects of a number of other factors also need to be investigated, including (1) shape complexity; (2) data noise; and (3) the aggregation methods. Methods to gain more sensitivity by perturbation, as briefly mentioned in Fig. 6, may also be studied. It would also be interesting to consider aggregating the results from different registration methods.

Acknowledgments: Thanks to the anonymous reviewers for their comments. This work was supported by the NIH under grants R01EB006494 (XP), R01MH070902 (HPB), R01MH069747 (HPB), R01NS35193 (LHS, XP), and the National Alliance for Research in Schizophrenia and Depression (FW, HPB).

References

[1] D. Adams, F. Rohlf, and D. Slice. Geometric morphometrics: ten years of progress following the revolution. *Italian Journal of Zoology*, 71(1):5–16, 2004.

[2] J. Ashburner and K. Friston. Voxel-Based MorphometryThe Methods. *Neuroimage*, 11(6):805–821, 2000.

[3] R. Blackith and R. Reyment. *Multivariate Morphometrics*. London, 1971.

[4] L. Breiman. Bagging Predictors. *Machine Learning*, 24(2):123–140, 1996.

[5] H. Chui, J. Zhang, and A. Rangarajan. Unsupervised learning of an atlas from unlabeled point-sets. *IEEE Trans. Patt. Anal. Mach. Intell*, 26:160–173, 2004.

[6] M. Chung, K. Dalton, L. Shen, A. Evans, and R. Davidson. Weighted Fourier Series Representation and Its Application to Quantifying the Amount of Gray Matter. *Medical Imaging, IEEE Transactions on*, 26(4):566, 2007.

[7] R. Davies. *Learning Shape: Optimal Models for Analysing Shape Variability*. PhD thesis, University of Manchester, 2002.

[8] T. Dietterich. Ensemble Methods in Machine Learning. In *Multiple Classifier Systems*, pages 1–15. Springer, 2000.

[9] I. L. Dryden and K. V. Mardia. *Statistical Shape Analysis*. John Wiley and Sons, 1998.

[10] J. M. Fitzpatrik. *Medical Image Registration*, chapter Detection failure, assessing success, page 117139. CRC Press, Baton Rouge, Florida, 2001.

[11] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *ICML*, pages 148–156, 1996.

[12] P. Golland and B. Fischl. Permutation tests for classification: Towards statistical significance in image-based studies. In *IPMI*, pages 330–341, 2003.

[13] P. Golland, W. Grimson, M. Shenton, and R. Kikinis. Detection and analysis of statistical differences in anatomical shape. *Medical Image Analysis*, 9(1):69–86, 2005.

[14] T. Heimann, I. Oguz, I. Wolf, M. Styner, and H. Meinzer. Implementing the Automatic Generation of 3D Statistical Shape Models with ITK. In *Open Science Workshop at MICCAI, Copenhagen*, 2006.

[15] P. Hellier, C. Barillot, I. Corouge, B. Gibaud, G. Le Goualher, D. Collins, A. Evans, G. Malandain, N. Ayache, G. Christensen, et al. Retrospective evaluation of intersubject brain registration. *Medical Imaging, IEEE Transactions on*, 22(9):1120–1130, 2003.

[16] X. Huang, N. Paragios, and D. N. Metaxas. Shape registration in implicit spaces using information theory and free form deformations. *IEEE Trans. Patt. Anal. Mach. Intell.*, 28(8):1303–1318, 2006.

[17] M. Styner, I. Oguz, S. Xu, C. Brechbuhler, D. Pantazis, J. Levitt, M. Shenton, and G. Gerig. Framework for the statistical shape analysis of brain structures using spharm-pdm. In *Open Science Workshop at MICCAI*, 2006.

[18] H. H. Thodberg. Minimum description length shape and appearance models. In *Proc. IPMI*, pages 51–62. BMVA, 2003.

[19] S. J. Timoner, P. Golland, R. Kikinis, M. E. Shenton, W. E. L. Grimson, and W. M. W. III. Performance issues in shape classification. In *MICCAI (1)*, pages 355–362, 2002.

[20] P. Yu, P. Grant, Y. Qi, X. Han, F. Segonne, R. Pienaar, E. Busa, J. Pacheco, N. Makris, R. Buckner, et al. Cortical Surface Shape Analysis Based on Spherical Wavelets. *Medical Imaging, IEEE Transactions on*, 26(4):582, 2007.

[21] L. Zhou, R. Hartley, P. Lieby, N. Barnes, K. Anstey, N. Cherbuin, and P. Sachdev. A study of hippocampal shape difference between genders by efficient hypothesis test and discriminative deformation. In *MICCAI (1)*, pages 375–383, 2007.