Multiatlas Segmentation as Nonparametric Regression

Suyash P. Awate and Ross T. Whitaker, *Fellow, IEEE*

***Abstract*—This paper proposes a novel theoretical framework to model and analyze the statistical characteristics of a wide range of segmentation methods that incorporate a database of label maps or atlases; such methods are termed as label fusion or *multiatlas segmentation*. We model these multiatlas segmentation problems as *nonparametric regression* problems in the high-dimensional space of image patches. We analyze the nonparametric estimator's convergence behavior that characterizes *expected segmentation error* as a function of the *size* of the multiatlas database. We show that this error has an analytic form involving several parameters that are fundamental to the specific segmentation problem (determined by the chosen anatomical structure, imaging modality, registration algorithm, and label-fusion algorithm). We describe how to estimate these parameters and show that several human anatomical structures exhibit the trends modeled analytically. We use these parameter estimates to *optimize* the regression estimator. We show that the expected error for large database sizes is well *predicted* by models learned on small databases. Thus, a few expert segmentations can help predict the database sizes required to keep the expected error below a specified tolerance level. Such cost-benefit analysis is crucial for deploying clinical multiatlas segmentation systems.**

***Index Terms*—k-nearest-neighbor (kNN), label fusion, multiatlas, nonparametric, regression, segmentation.**

I. Introduction

The strategy of segmenting an image using available segmentations of similar images, termed "segmentation by example," has lead to various approaches in a wide spectrum of biomedical applications over the last two decades [1], [2]. This paper deals with segmentation methods [3]–[5] that use a combination of 1) a set of *template* images that depict anatomical structures and 2) for each template, a set of one or more tissue probability maps or label maps or *segmentation* maps that give the probability of each voxel belonging to a specific anatomical structure. A pair of images comprising a template image and an associated segmentation map is termed an *atlas*. Because these methods employ multiple atlases, they are termed as *multiatlas* segmentation methods. In this paper, references to *a specific segmentation problem* mean a biomedical image segmentation problem that is determined by the choice of the anatomical structure, imaging modality, registration algorithm, and label-fusion algorithm.

Atlas-based segmentation is most relevant for segmenting such anatomical structures whose boundary parts fail to be readily detectable in the image data alone. For instance, in T1-weighted magnetic resonance (MR) brain images, subcortical brain structures have boundary parts with very low contrast-to-noise ratios between regions on either side of the boundary. With limited information present in single-voxel intensity, atlas-based segmentation relies heavily on the information present in a large spatial neighborhood of the voxel.

Traditional atlas-based methods, deforming a single presegmented template to match the *target*, leverage information within the spatial configuration of those surrounding structures that have well defined boundaries in the images. This relies on the well founded biological assumption that the geometry (i.e., location, pose, size, and shape) of the weakly visible structure is a (stochastic) function of the geometry of these surrounding structures. With a usually reliable matching of these surrounding structures, the registration gives a deformation that best matches the weakly visible structure. This deformation applied to the template segmentation gives a target segmentation.

In recent years, large collections of medical images, and associated expert-defined segmentations, are becoming ubiquitous as open resources, and within specific clinical practices. This has lead to multiatlas or label-fusion segmentation methods [3]–[12] that leverage information in the entire *database* of atlases. Given a large database, multiatlas approaches, in practice, first select a small subset of templates that are most similar to the target. They independently register the selected templates to the target and, then, deform database segmentations to the target space. A weighted average of the deformed segmentations produces an estimate of the segmentation of the target. Instead of using the entire database, the carefully selected subset produces better results, as shown for brain [3], [5] and cardiac [4] MR images. Similarly, well tuned weighting schemes [8], [12] produce better results. Note that selecting a subset of templates is equivalent to assigning zero weights to all other templates. The proposed theoretical framework and the associated results shed

Manuscript received March 09, 2014; revised April 27, 2014; accepted April 27, 2014. Date of publication April 30, 2014; date of current version August 28, 2014. The authors gratefully acknowledge the support of this work through the National Alliance for Medical Image Computing U54-EB005149 and the NIH/NCRR Center for Integrative Biomedical Computing P41-RR12553. AtlasWerks is an open source software project that is principally funded through NIH/NIBIB 5R01EB007688 and NIH/NCRR 2P41 RR0112553-12. The OAI is a public-private partnership comprised of five contracts (N01-AR-2-2258; N01-AR-2-2259; N01-AR-2-2260; N01-AR-2-2261; N01-AR-2-2262) funded by the NIH, a branch of the Department of Health and Human Services, and conducted by the OAI Study Investigators. Private funding partners include Merck Research Laboratories; Novartis Pharmaceuticals Corporation, GlaxoSmithKline; and Pfizer, Inc. *Asterisk indicates corresponding author.*

*S. P. Awate is with the Computer Science and Engineering Department, Indian Institute of Technology (IIT) Bombay, Mumbai 400076, India (e-mail: suyash@cs.utah.edu).

R. T. Whitaker is with the Scientific Computing and Imaging (SCI) Institute and the School of Computing, University of Utah, Salt Lake City, UT 84112 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMI.2014.2321281

light on this behavior, indicating that an optimal subset size, or the weighting scheme, depends on the database size.

This paper makes several contributions. It proposes a novel statistical nonparametric regression framework to model a class of multiatlas segmentation approaches and analyze the *convergence* behavior of expected segmentation error with respect to database size. It shows that the expected segmentation error has a specific analytic form involving several parameters that are fundamental to the specific segmentation problem. By measuring these parameters, it characterizes the specific segmentation problem and method in terms of 11) the complexity of the function mapping the geometry of (clearly visible) surrounding structures to the geometry of the structure of interest, 2) the complexity of the function mapping local template appearance to the segmentation, 3) the inherent anatomical randomness in the structure's geometry, 4) number of atlases available in the database, and 5) the label-fusion weighting scheme. Furthermore, we use these parameter estimates to further optimize the regression estimator. In this way, the framework offers new methods to evaluate the efficacy of a particular database of atlases, imaging modality, registration algorithm, and label-fusion algorithm. We show that the expected error for large database sizes is well *predicted* by models learned on small databases. Thus, a few expert segmentations can help predict the database sizes needed to keep the expected error below a specified tolerance level. Such cost-benefit analysis is crucial for deploying clinical decision support systems involving multiatlas segmentation.

II. Related Work

This section describes the relationships between the literature on atlas-based segmentation and the proposed framework.

There exists a large body of recent work proposing many variations of multiatlas segmentation methods. For instance, some recent approaches for label fusion have found improvements in performance by using locally weighted averaging where the tissue probability at a voxel is determined by using only that information in the (registered) atlases which lies within the locality of that voxel [6], [8], [10]. Other approaches have found that generalized weighting schemes [8], [12] perform better. Results in [3]–[5], [7] show that a careful selection of templates (also a kind of a weighting scheme), e.g., selecting the top-few most-similar templates for target segmentation, performs best. This is consistent with the results in this paper that indicate the existence of an optimal number of atlases to use for a given database size. This paper further shows that this optimal number increases with the size of the multiatlas database. The proposed multiatlas segmentation approach also incorporates such a strategy by defining weighted-average schemes 1) separately at each voxel, relying on local similarities between the target and the templates, and 2) that select a small number of templates dependent on database size. The focus of the proposed framework is less on the specific similarity measure or the specific weighting function used, but more on the characterization of the expected segmentation error, as a function of database size, for any given local similarity measure and weighting function. Indeed, the proposed theoretical formulation is general enough to allow for modeling and analysis of any such scheme.

This manuscript offers significant improvements over our previous work [13] as follows. First, this paper formulates multiatlas segmentation using a combination of separate local regression functions, relying on local template-similarity kernels, at every voxel in the image. Second, to evaluate the quality of prediction of the expected segmentation error for large databases using small databases, it presents confidence estimates in the form of the variation of the predictions over different instances of small databases available. Third, this paper describes a method for optimizing, as a function of database size, the parameter corresponding to the number of nearest neighbors/templates to use in the k-nearest-neighbor (kNN) nonparametric regression estimator. Fourth, in addition to the brain atlas database used in [13], this paper demonstrates the efficacy of the proposed modeling and prediction on a large knee MR atlas database for knee-cartilage segmentation.

A parametric model for the Dice similarity as a function of the 1) random and systematic errors resulting from, e.g., misregistration or atlas inconsistencies and 2) size of the atlas database appears in the pioneering work in [3], [14], [15]. Although [3] motivates the model primarily for quantifying segmentation errors for a given database size, the insightful model appears to be more general and has a wider utility. In this spirit, we use this model for predicting the Dice similarity using a large number of atlases, by learning the model using a small atlas database. The approach in [3] has some limitations that are overcome by our approach. Unlike our proposed approach, the model in [3] 1) focuses on the analysis of the entire structure because it uses Dice similarity as the performance measure and is, consequently, inapplicable to voxelwise analysis, 2) ignores the number of atlases, i.e., k, in the weighted average that gives the multiatlas segmentation, and thereby, would entail learning separate models of segmentation performance for every possible value of k, and 3) is unable to predict the optimal number templates to be used for any given database size—this will be impossible for [3] if the optimal k, for a certain large database size, is larger than the size of the training database used to learn the models.

Some multiatlas segmentation approaches deal with specific kinds of correlations in label maps or biomedical images. For instance, some methods focus on compensating for inter-voxel label correlations via second-order polynomial regression over small (3^3 or 5^3) image patches [11]. Others adapt to the correlations within a group of target images by proposing simultaneous registration between a group of target images and the group templates in the atlas database [9]. The spirit of this paper is quite different from that of the aforementioned approaches. This paper focuses on predicting expected segmentation error as a function of database size and by modeling the segmentation problem as a regression problem and estimating the regression parameters.

The image clustering method in [16] uses various consistency criteria to compute an optimal number of clusters to represent a given population of images. They perform label fusion using the entire set of cluster means, termed templates. Their approach is orthogonal to the one in this paper that finds the optimal subset of the template database to use for label fusion, as the template-database size grows.

Some early atlas-based segmentation approaches focus on estimating rater-performance parameters (particularly, rater bias) like STAPLE [17] and the parameters' confidence intervals [18], where multiple segmentations exist for a single biomedical image. In contrast, this manuscript focuses on segmentation strategies that combine one or more segmentations from multiple biomedical images and analyzes the number of atlases (somewhat analogous to raters) required to keep the expected error below a specified tolerance level.

III. Methods

This section presents a novel statistical framework, relying on *nonparametric regression* [19], to model and analyze a class of multiatlas segmentation approaches.

Consider the problem of estimating the unknown segmentation for a target image, using a database of atlases. In this context, 1) each atlas is a pair comprising one template and one probabilistic segmentation and 2) a single template could be a part of multiple atlases if it has multiple associated probabilistic segmentations where each probabilistic segmentation could have arisen from a group of binary expert segmentations or a single deformed binary segmentation.

Treating each atlas as a *member of a family of atlases* under constrained diffeomorphisms (e.g., constrained under limited deformation norm), we first transform the database to factor out a diffeomorphism between the geometrical configurations of anatomical structures within the target and each template; better matches of the two geometries would usually lead to better matches of the segmentations. We assume that multiatlas segmentation methods can compute an optimal diffeomorphism using image registration on the raw intensities or on derived geometry-capturing features and, later, deform each template and segmentation, in the database, to the target image's physical space. Thus, we propose to characterize the difficulty for a specific segmentation problem by 1) modeling multiatlas segmentation as a regression problem where the *independent variable* represents *deformed template images* and the *dependent variable* represents *deformed segmentation images* and 2) analyzing the convergence of the expected segmentation error with respect to increasing database sizes.

A. Multiatlas Segmentation as Nonparametric Regression

Let F be a vector random variable that models a biomedical image (diffeomorphically deformed to a common anatomical space; without any loss of generality) with V voxels. Note that the assumption of the images being warped to a common anatomical space is mainly theoretical in nature and is *not* practically restrictive because the warps are assumed to be diffeomorphic (i.e., smooth and invertible); thus, theoretically, the analysis could be performed in the coordinate space of any target or template. The associated probability density function (PDF) $P(F)$ generates observed images $f \in \mathbb{R}^V$. For a specific anatomical structure in the image, let S be a V-dimensional vector random variable modeling the deformed segmentation map that is non-binary or probabilistic. The associated PDF $P(S)$ generates observed segmentations $s \in \mathbb{R}^V$. Let S_v denote the random variable at the vth component of S (i.e., voxel v in image). Then, $s_v \in [0,1]$. We assume that the joint PDF $P(F,S)$ captures dependencies between biomedical images f and their segmentations s.

Consider a *database* $a^M := \{(f^m, s^m)\}_{m=1}^M$ of M atlases, i.e., *template* images $\{f^m\}_{m=1}^M$ paired with their *true segmentations* $\{s^m\}_{m=1}^M$, where each observed image pair (f^m, s^m) is drawn independently from the PDF $P(F,S)$. For a given *target* image f^0 whose true segmentation s^0 is *unknown*, multiatlas segmentation methods use the database a^M to get an estimate $\widehat{s}^0$ of the true segmentation.

We model multiatlas segmentation as nonparametric regression. Let $r(F) : \mathbb{R}^V \mapsto \mathbb{R}^V$ be a regression function of the dependent variable S on the independent variable F. From the class of regression functions, we choose $r(F)$ as the regression function that minimizes the mean squared error risk function $E_{P(F,S)}\left[\|S - r(F)\|^2\right] = E_{P(F)}\left[E_{P(S|F)}\left[\|S - r(F)\|^2\right]\right]$. For any target f, the risk-minimizing regression function is the conditional expectation $r(f) := E_{P(S|f)}[S]$. Let $\widehat{r}(F, a^M)$ be an estimator of the true conditional expectation $r(F)$, which relies on the atlas database a^M.

For a specific segmentation problem, we want to characterize the behavior of a conditional-expectation regression estimator over varying images f and varying databases a^M. Hence, we treat the database as a random variable $\mathcal{A}^M := \langle F^1, S^1, \ldots, F^M, S^M \rangle$ and then define a joint PDF $P(F, S, \mathcal{A}^M)$ and a new mean-squared-error (MSE) function as follows. We define the joint PDF $P(F, S, \mathcal{A}^M) := P(F,S)P(\mathcal{A}^M)$, assuming independence of the observed image pair (f, s) and the database $\mathcal{A}^M$. We define $P(\mathcal{A}^M) := \Pi_{m=1}^M P(F^m, S^m)$, assuming that each atlas is generated independently from the same distribution $P(F,S)$ that generates target images and their segmentations. To capture the *expected segmentation error*, we define the MSE function

$$\mathcal{E}(M) := E_{P(F,S,\mathcal{A}^M)}\left[\left\|S - \widehat{r}(F, \mathcal{A}^M)\right\|_2^2\right]. \quad (1)$$

We want to model the regression functions for the entire image using a combination of separate (local) regression functions at every voxel v in the image. With this motivation, let $r_v(f)$ and $\widehat{r}_v(f, \mathcal{A}^M)$ denote the vth components of regression functions $r(f)$ and $\widehat{r}(f, \mathcal{A}^M)$, respectively, which correspond to the regression functions at the vth voxel in the image that produce the probabilistic segmentation at the vth voxel in the image. Associated with the regression functions at voxel v in the image is the MSE at voxel v in the image, denoted by $\mathcal{E}_v(M)$. Then, the linearity of expectation gives

$$\mathcal{E}(M) = \sum_{v=1}^{V} \mathcal{E}_v(M), \text{where} \quad (2)$$

$$\mathcal{E}_v(M) := E_{P(F,S,\mathcal{A}^M)}\left[\left(S_v - \widehat{r}_v(F, \mathcal{A}^M)\right)^2\right]. \quad (3)$$

Further analysis gives

$$\mathcal{E}_v(M) = E_{P(F)}\left[\mathcal{E}_v(M, F)\right], \text{where} \quad (4)$$

$$\begin{aligned}\mathcal{E}_v(M, f) &:= E_{P(S,\mathcal{A}^M|f)}\left[\left(S_v - \widehat{r}_v(f, \mathcal{A}^M)\right)^2\right] \quad (5)\\ &= E_{P(S|f)}\left[\left(S_v - r_v(f)\right)^2\right]\\ &\quad + E_{P(\mathcal{A}^M|f)}\left[\left(r_v(f) - \widehat{r}_v(f, \mathcal{A}^M)\right)^2\right]\\ &\quad + E_{P(S,\mathcal{A}^M|f)}\Big[2\left(S_v - r_v(f)\right)\\ &\qquad \cdot \left(r_v(f) - \widehat{r}_v(f, \mathcal{A}^M)\right)\Big]. \quad (6)\end{aligned}$$

The second term in the expansion of the expression $\mathcal{E}_v(M,f)$ leads to $E_{P(F)}E_{P(\mathcal{A}^M|F)}\left[(r_v(F)-\widehat{r}_v(F,\mathcal{A}^M))^2\right]$, termed the mean integrated squared error [19]. We now analyze all three terms in the expression for $\mathcal{E}_v(M,f)$.

1) For the conditional-expectation regression function $r_v(f) := E_{P(S|f)}[S_v]$, the first term is the variance of the conditional PDF $P(S_v|f)$, i.e.,

$$\mathcal{V}(S_v|f) := E_{P(S|f)}[(S_v - r_v(f))^2]. \tag{7}$$

This term 1) depends on the inherent (beyond our control) randomness in the segmentation, given image data f and 2) is independent of the estimator $\widehat{r}_v(f,\mathcal{A}^M)$.

2) The second term relates to the quality of approximation, of the estimator $\widehat{r}_v(f,\mathcal{A}^M)$, to the true conditional-expectation regression function $r_v(f)$. This term equals the sum of the estimator's squared bias and the estimator's variance. This term depends on 1) the database size M and 2) the characteristics of the marginal distribution $P(F)$ and the regression function $r_v(\cdot)$ in the locality of f.

Note that the estimator bias is the difference between the true conditional expectation and the expected value of the estimator, i.e.,

$$\mathcal{B}(\widehat{r}_v(f,\mathcal{A}^M)) := r_v(f) - E_{P(\mathcal{A}^M|f)}[\widehat{r}_v(f,\mathcal{A}^M)]. \tag{8}$$

The estimator variance is the expected value of the squared difference between 1) the estimator and 2) the expected value of the estimator, i.e.,

$$\mathcal{V}(\widehat{r}_v(f,\mathcal{A}^M)) \\ := E_{P(\mathcal{A}^M|f)}[(\widehat{r}_v(f,\mathcal{A}^M) - E_{P(\mathcal{A}^M|f)}[\widehat{r}_v(f,\mathcal{A}^M)])^2]. \tag{9}$$

3) The third term vanishes because it is equal to $E_{P(\mathcal{A}^M)}E_{P(S|\mathcal{A}^M,f)}[2\left(S_v - r_v(f)\right)\left(r_v(f)-\widehat{r}_v(f,\mathcal{A}^M)\right)]$ where the inner expectation is zero because 1) $S_v - r_v(f)$ and $r_v(f) - \widehat{r}_v(f,\mathcal{A}^M)$ form a decomposition of the random variable $S_v - \widehat{r}_v(F,\mathcal{A}^M)$ and 2) $E_{P(S|f)}\left[S_v - r_v(f)\right] = 0$.

Thus, $\mathcal{E}_v(M,f)$ equals the sum of the variance of the conditional PDF, the squared bias of the estimator, and the variance of the estimator, i.e.,

$$\mathcal{E}_v(M,f) = \mathcal{V}(S_v|f) + \mathcal{B}^2(\widehat{r}_v(f,\mathcal{A}^M)) + \mathcal{V}(\widehat{r}_v(f,\mathcal{A}^M)). \tag{10}$$

B. Multiatlas Segmentation as Local Generalized-kNN Regressions in Kernel Feature Spaces

We now choose a specific regression estimator. A consistent estimator for the conditional-expectation regression function $r_v(f)$ is the *generalized-kNN* estimator [20]

$$\widehat{r}_v(f,a^M) := \frac{\sum_{m=1}^{M} s_v^m w\left(\frac{\Phi_v(f^m)-\Phi_v(f)}{R_k}\right)}{\sum_{m=1}^{M} w\left(\frac{\Phi_v(f^m)-\Phi_v(f)}{R_k}\right)} \tag{11}$$

where $\Phi_v(\cdot) : \mathbb{R}^V \mapsto \mathcal{H}$ is the *feature map* associated with a Mercer kernel that maps images to a Hilbert space $\mathcal{H}$, R_k is the $\|\cdot\|_{\mathcal{H}}$-normed distance between $\Phi_v(f)$ and its kth nearest neighbor in the set $\{\Phi_v(f^m)\}_{m=1}^{M}$, and $w(\cdot) : \mathcal{H} \to \mathbb{R}$ is a bounded non-negative generalized weighting function satisfying $\int_{\mathcal{H}} w(u)du = 1$ and $w(u) = 0, \forall u : \|u\|_{\mathcal{H}} > 1$.

For the class of generalized-kNN estimators [20], the bias $\mathcal{B}$ and variance $\mathcal{V}$ are

$$\mathcal{B}(\widehat{r}_v(f,\mathcal{A}^M)) \doteq \varphi\Big(r_v(\cdot),\Phi_v(\cdot),P(F),f,d_v\Big)\left(\frac{k}{M}\right)^{2/d_v} \tag{12}$$

and

$$\mathcal{V}(\widehat{r}_v(f,\mathcal{A}^M)) \doteq \psi\Big(\Phi_v(\cdot),w(\cdot),d_v\Big)\frac{\mathcal{V}\left(S_v|f\right)}{k} \tag{13}$$

where 1) d_v is the dimension of the mapped independent variable $\Phi_v(F)$ in $\mathcal{H}$; 2) $\varphi\big(r_v(\cdot),\Phi_v(\cdot),P(F),f,d_v\big)$ depends on feature map $\Phi_v(\cdot)$, the values and differential properties of the PDF $P(F)$ in the locality of the fixed image f, the local differential properties of the true regression function $r_v(\cdot)$, and dimension d_v; 3) $\psi\big(\Phi_v(\cdot),w(\cdot),d_v,\big)$ depends on the chosen weight function $w(\cdot)$ and the dimension d_v. Indeed, the kNN estimator converges to the true conditional-expectation regression function asymptotically as the database size $M \to \infty$ and the number of nearest neighbors is chosen (dependent on the database size M) such that $k \to \infty$ at an appropriate rate such that $k/M \to 0$.

It is important to note that the rate of convergence of the bias and variance depends on 1) the dimensionality d_v associated with the independent random variable F, 2) the values and the differential properties of the PDF $P(F)$ of images, and 3) the differential properties of the regression function $r_v(f)$.

1) Choice of the Feature Map $\Phi_v(\cdot)$: The proposed generalized-kNN framework gives us the flexibility to choose or design a distance metric on the space of biomedical images through the choice of a Mercer kernel or the associated feature map on images. In this way, images can be first mapped to a high-dimensional Hilbert space where distances can be evaluated through kernel evaluation. In this paper, we use a simple Mercer kernel, i.e., the local normalized cross correlation $\mathcal{K}_v(\cdot,\cdot)$ to measure local image similarity. Specifically, for the regression estimator $\widehat{r}_v(f,a^M)$ at voxel v, we define the distance $d_v(f,f')$ between two images f, f' by using image patches f_v, f'_v centered at v as

$$d_v^2(f,f') = \mathcal{K}_v(f,f) + \mathcal{K}_v(f',f') - 2\mathcal{K}_v(f,f') \tag{14}$$

where $\mathcal{K}_v(\cdot,\cdot)$ is the normalized Mercer kernel

$$\mathcal{K}_v(f,f') := \frac{\widetilde{\mathcal{K}}_v(f,f')}{\sqrt{\widetilde{\mathcal{K}}_v(f,f)\widetilde{\mathcal{K}}_v(f',f')}}, \text{ where} \tag{15}$$

$$\widetilde{\mathcal{K}}_v(f,f') := \langle \breve{f}_v, \breve{f}'_v \rangle \tag{16}$$

and $\breve{f}_v$ is the mean-subtracted version of patch f_v. The Appendix gives a proof of $\mathcal{K}_v(\cdot,\cdot)$ being a Mercer kernel. With this kernel, at voxel v, the feature map is $\Phi_v(\cdot) : f \mapsto \breve{f}_v / \left\|\breve{f}_v\right\|$, the inner product for $h, h' \in \mathcal{H}$ is $\langle h, h' \rangle_{\mathcal{H}} = \mathcal{K}_v(f,f')$, the squared norm is $\|h\|^2_{\mathcal{H}} = \langle h, h\rangle_{\mathcal{H}}$, and the local squared distance between images f and f' is $d_v^2(f, f') = 2(1 - \langle \breve{f}_v, \breve{f}'_v\rangle)$.

In practice, the mean subtraction and rescaling help in locally standardizing the intensities in MR images, while effectively addressing local variability in image contrast, intensity nonuniformity, and intensity scale in MR imaging. The experiments in this paper use an isotropic patch of size 21^3 mm^3. The proposed framework indeed allows the usage of more sophisticated Mercer kernels such as the pyramid match kernel [21] and the spatial pyramid kernel [22] that was used for multiatlas segmentation in [13], [23].

2) Choice of the Weighting Function $w(\cdot)$: Given that the mapped data $\Phi_v(f)$ lie in a Hilbert space, $\psi\big(\Phi_v(\cdot), w(\cdot), d_v\big) = c(d_v)\int w^2(u)du$, where $c(d_v)$ is the volume of the unit sphere in d_v dimensions of $\mathcal{H}$ [20]. For simplicity, this paper chooses $w(u)$ to be constant $\forall u : \|u\|_{\mathcal{H}} \leq 1$, which leads to $\psi\big(\Phi_v(\cdot), w(\cdot), d_v\big) = 1$.

C. Parametric Model for Expected Segmentation Error $\mathcal{E}(M)$ as a Function of Database Size M

For the chosen kNN scheme, at each voxel v, the expected segmentation error is parametrized as

$$\mathcal{E}_v(M, k) \doteq \alpha_v\left(1 + \frac{1}{k}\right) + \beta_v\left(\frac{k}{M}\right)^{4/d_v} \tag{17}$$

$$\text{where } \alpha_v = E_{P(F)}\left[\mathcal{V}(S_v|F)\right] \tag{18}$$

$$\text{and } \beta_v = E_{P(F)}\left[\varphi^2\left(r_v(\cdot), \Phi_v(\cdot), P(F), F, d_v\right)\right]. \tag{19}$$

In practice, the success of the model fitting, in Section IV, indicates that β_v is relatively independent of d_v. The expected segmentation error for the entire anatomical structure is

$$\begin{aligned}\mathcal{E}(M, k) &= \sum_{v=1}^{V} \mathcal{E}_v(M, k)\\ &= \left(1 + \frac{1}{k}\right)\sum_{v=1}^{V}\alpha_v + \sum_{v=1}^{V}\beta_v\left(\frac{k}{M}\right)^{4/d_v}.\end{aligned} \tag{20}$$

These equations capture the voxelwise characteristics of a specific segmentation problem and approach through parameters α_v, β_v, d_v, whose significance we describe next.

- α_v is the integrated conditional variance of the segmentations and denotes the intrinsic randomness in the segmentations s_v at voxel v as a function of the image data f. α_v is independent of the regression estimator and hence is the lowest possible achievable MSE $\mathcal{E}(M)$ at voxel v for the specific segmentation problem. For the chosen kNN regression estimator, this lowest MSE is achieved when the regression estimator converges to the true conditional expectation. As $M \to \infty$, we can make the kNN estimator converge to the conditional expectation, by letting k tend to ∞ at such a rate so that $(k/M) \to 0$.
- β_v represents the overall complexity of multiatlas segmentation in terms of the 1) differential properties of the true regression function $r_v(f)$ and 2) values and differential properties of the image PDF $P(F)$. $r_v(\cdot)$ is harder to estimate when β_v is increased because: 1) larger gradients and curvatures in $r_v(\cdot)$ lead to larger values of φ; 2) around a target f_0, low values of $P(F)$ make it harder to obtain databases comprising sufficiently many templates near f_0; 3) around a target f_0, locally varying $P(F)$ leads to databases where the templates near f_0 pull the segmentation estimate towards that for the local higher-probability templates.
- When the class of signals F is unconstrained, d_v equals the number of voxels in the image patch minus 2 (by enforcing zero mean and unit norm), which can be quite large, i.e., several hundreds or thousands. However, consistent with empirical evidence in the signal-processing and machine-learning literature that the intrinsic dimension of real-world multivariate data is far less than the number of variables used for representation [24]–[28], we consider d_v as the *intrinsic dimension* of the independent variable (feature-space-mapped template-image patches) at voxel v. Larger d_v increases the difficulty of multiatlas segmentation by requiring estimation of a higher-dimensional regressor.

D. Fitting the Parametric Model for Expected Segmentation Error $\mathcal{E}(M)$

This section builds upon the theory described in previous sections to estimate the set of parameters $\{\alpha_v, \beta_v, d_v\}_{v=1}^{V}$ for all voxels and to subsequently estimate an optimal k as a function of the database size M.

1) Empirical Computation of the Voxelwise Expected Segmentation Error $\mathcal{E}_v(M)$: Consider an atlas database $b^N = \{g^n, t^n\}_{n=1}^{N}$ with N atlases available for analysis. For a chosen number of nearest neighbors k and a chosen database size $M \leq N$, we propose to empirically compute $\mathcal{E}_v(M, k)$ in (3), for all voxels v, by: 1) Monte-Carlo bootstrap sampling of target images $f \in \{g^n\}_{n=1}^{N}$ to evaluate the expectation over $P(F)$, 2) for each f, Monte-Carlo bootstrap sampling of segmentations $s \in \{t^n\}_{n=1}^{N}$ to evaluate the expectation over $P(S|F)$ 3) for each pair (f, s), Monte-Carlo bootstrap sampling of databases $a^M \subset b^N$ to evaluate the expectation $E_{P(\mathcal{A}^M|S,F)}[\cdot]$, and 4) computing the MSE value $\mathcal{E}_v(M, k)$, given the target segmentations s associated with the sampled f, at each voxel v. In this paper, 1) we sample 20 target images f, 2) the available databases give us only one probabilistic segmentation s for each f and, 3) for each (f, s) pair, we sample 20 databases a^M. Thus, we sample 400 pairs of (f, s, a^M). For the experiments in this paper, we found that a jackknife strategy for sampling f, s, a^M was also effective; indeed, jackknife estimation is known to be a linear approximation to bootstrap estimation [29].

We compute $\mathcal{E}_v(M, k)$ for a range of chosen values for database sizes $\{M_i\}_{i=1}^{I}$ and the chosen number of nearest neighbors $\{k_j\}_{j=1}^{J}$. This gives us the MSE values $\mathcal{E}_v(M_i, k_j)$. In this paper, we use the database sizes $M_1 = 10$, $\forall i > 1$, $M_i = M_{i-1} + 5$, and values $k_j \in \{5, 6, 7, 8, 9, 10\}$.

To estimate the variance of the subsequent parameter estimation with respect to the specific choice of the available database b^N, we perform Monte-Carlo bootstrap sampling of b^N to yield L databases $\{b_l^N\}_{l=1}^{L}$ and repeat the MSE $\mathcal{E}_v(M_i, k_j)$ calculation and the subsequent parameter estimation for each b_l^N. In this paper, we use $L = 20$.

2) Voxelwise Estimation of Parameters $\{\alpha_v, \beta_v, d_v\}$ With Spatial Regularization: Given the empirically computed MSEs $\mathcal{E}_v(M_i, k_j)$, for each bootstrap sample b_l^N, the estimation of parameters $\{\alpha_v, \beta_v, d_v\}_{v=1}^V$ (for a specific segmentation problem) leads to a weighted nonlinear least-squares curve-fitting problem at each voxel v, i.e.,

$$\min_{\alpha_v, \beta_v, d_v} \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{1}{\sigma_{vij}^2} \left(\mathcal{E}_v(M_i, k_j) - \alpha_v \left(1 + \frac{1}{k}\right) - \beta_v \left(\frac{k_j}{M_i}\right)^{4/d_v} \right)^2 \tag{21}$$

where weights $1/\sigma_{vij}^2$ are the inverse of the computed variances (over the L large-database samples b_l^N) associated with the set of squared errors $(S_v - \widehat{r}_v(f, \mathcal{A}^{M_i}))^2$ that produced the MSE $\mathcal{E}_v(M_i, k_j)$ for each M_i, k_j. The weights $1/\sigma_{vij}^2$ provide practical confidence measures for the Monte-Carlo estimates of the MSE $\mathcal{E}_v(M_i, k_j)$.

We realize that the characteristics of the segmentation problem at a voxel v, which is parameterized by α_v, β_v, d_v, will typically bear significant similarity to those at a nearby voxel v'. Thus, we reformulate this parameter estimation problem by incorporating a spatial regularization prior on the parameter values by modeling the image of parameters as a Markov random field (MRF). This spatial regularization also helps in avoiding local minima for the nonlinear fitting problem. For spatial regularization, we employ an MRF with a 26-neighbor system, use the set of two-voxel cliques $\mathcal{C}$, and the clique potential as the squared difference between the neighboring parameter values. Thus, the optimization problem becomes

$$\min_{\{\alpha_v, \beta_v, d_v\}_{v=1}^V} \sum_{v=1}^{V} \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{1}{\sigma_{vij}^2} \left(\mathcal{E}_v(M_i, k_j) - \alpha_v \left(1 + \frac{1}{k}\right) - \beta_v \left(\frac{k_j}{M_i}\right)^{4/d_v} \right)^2 + \lambda \sum_{(v,v') \in \mathcal{C}} \left((\alpha_v - \alpha_{v'})^2 + (\beta_v - \beta_{v'})^2 + (d_v - d_{v'})^2 \right). \tag{22}$$

We solve this optimization problem using a gradient-descent algorithm (with adaptive step size) that iteratively scans over the entire set of voxels and, within each scan, at each voxel v, performs iterative alternating minimization of the parameters α_v, β_v, d_v. Typically, very few scans are required for convergence of this gradient-descent strategy. In this paper, we simply use a single λ value for regularizing all three kinds of parameter values. We tune the regularization parameter λ empirically using the standard L-curve approach [30].

We perform the parameter estimation once for every Monte-Carlo bootstrap sample b_l^N of the available database. This gives L parameter estimates at each voxel, one resulting from each b_l^N, and, in turn, a mean value of each of the parameter estimates for α_v, β_v, d_v and the standard deviation, under variations in the particular database available for analysis.

E. Optimizing the Regression Estimator: Optimal Number of Nearest Templates $k^*(M)$ as a Function of Database Size M

The MSE $\mathcal{E}_v(M, k)$ at every voxel v for a chosen database size M depends on the number of nearest neighbors k used in the kNN regression. As described in previous sections, to achieve lowest possible MSE $\mathcal{E}(M)$ through kNN, k needs to be optimized as a function of M. Theoretically, we can choose an optimal k, at each voxel v, for each regression estimator $\widehat{r}_v(f, \mathcal{A}^M)$. However, voxel-specific optimal k values are impractical for a real-world application of multiatlas segmentation where a more suitable strategy would be to select a single optimal k, as a function of M, for the entire image.

After estimating the set of parameters $\{\alpha_v, \beta_v, d_v\}_{v=1}^V$ over all voxels, we optimize k for a specific segmentation problem and any given database size M by minimizing $\mathcal{E}(M, k)$ using a gradient-descent algorithm with adaptive step size. This gives us the optimal number $k^*(M)$ of nearest neighbors/templates to use for any given database size M.

We perform the estimation of the optimal number of nearest neighbors/templates $k^*(M)$ once for every Monte-Carlo bootstrap sample b_l^N of the available database. This provides us L values of parameter estimates at each voxel, one resulting from each b_l^N. This gives us, at each voxel, a mean value of the parameter estimate $k^*(M)$ and its standard deviation, under variations in the particular database available for analysis.

F. Predicting Expected Segmentation Error $\mathcal{E}(M)$ for Large Database Sizes M

One of the motivations for characterizing the difficulty of a specific segmentation problem, using the parameter estimation, is to be able to predict the expected segmentation error for database sizes much larger than those available for analysis. To demonstrate this aspect, we perform Monte-Carlo bootstrap sampling of smaller databases $b^{\widetilde{N}}$ of size $\widetilde{N} \ll N$ and assume that we only had a database of size $\widetilde{N}$ for analysis. In this paper, we use $\widetilde{N} = 41$ that is roughly four times smaller than N. Subsequently, we perform parameter estimation for each Monte-Carlo bootstrap-sampled database in $\{b_l^{\widetilde{N}}\}_{l=1}^L$. Having characterized a specific segmentation problem by estimating the parameters $\{\alpha_v, \beta_v, d_v\}_{v=1}^V$, for any given database $b_l^{\widetilde{N}}$, we can find the optimal k, i.e., $k^*(M)$, for any $M \gg \widetilde{N}$ and subsequently predict the expected segmentation error $\mathcal{E}(M, k^*(M))$. The bootstrap analysis provides us with L estimates of the expected segmentation error, thus informing us about the variability in the estimation process over stochastic variations in the database available for analysis. In this way, we can predict the database size needed to keep the expected segmentation error below a specified tolerance level. In this paper, we *predict* the expected segmentation error for a range of database sizes M, where $\widetilde{N} < M \leq N$ and *validate* the quality of the prediction using the expected segmentation error values computed from the entire available database of size N as described in previous sections.

IV. RESULTS

This section describes some practical considerations and shows results on two large clinical databases. The results demonstrate the validity of the proposed model for multiatlas segmentation and the utility of the proposed analysis in clinical applications. Section IV-D shows that several human anatomical structures exhibit the parametric trends determined by the model, thereby showing that the model is well suited for real-world applications. Section IV-E shows results of predicting the expected segmentation error for database sizes larger than those available for analysis. It validates the quality of the prediction using the results in Section IV-D. In this way, small databases (requiring few expert segmentations) can be used to predict the database sizes required to keep the expected segmentation error below a specified tolerance level.

In this paper, we obtain the multiatlas segmentation at each voxel v, using a kNN estimator $\widehat{r}_v(\cdot,\cdot)$ where the weighting function $w(\cdot)$ is constant for all the k nearest neighbors. For such an estimator at voxel v, we measure distances between images via the local normalized cross correlation, a Mercer kernel, using isotropic $21^3\ \mathrm{mm}^3$ image patches around voxels v. At each voxel, we fit parametric curves to the empirically computed MSE as a function of database size, incorporating spatial regularization of the parameter estimates via a parameter λ that we tune empirically using the standard L-curve approach [30].

A. Clinical Databases

We use two large clinical databases b^N for evaluation.

1) From the National Alliance for Medical Image Computing (NAMIC; www.na-mic.org), we obtain a database comprising $N = 186$ T1-weighted MR brain images (dimensions $\approx 256 \times 256 \times 240$; voxels $\approx 1\ \mathrm{mm}^3$ isotropic) with expert segmentations for the caudate, putamen, thalamus, hippocampus, and globus pallidus in both hemispheres. In this paper, we combine the pair of corresponding structures in the two brain hemispheres in a single analysis.
2) From the Osteoarthritis Initiative (OAI; www.oai.ucsf.edu), we obtain a publicly available database comprising $N = 140$ T1-weighted MR knee images (dimensions $\approx 230 \times 150 \times 270$; voxels $\approx 1\ \mathrm{mm}^3$ isotropic) with expert segmentations for the meniscus, patellar, and tibial cartilages.

Both databases provide only one segmentation t for each anatomical structure within each MR image g.

To compute expected errors for multiatlas segmentation using an atlas database of size M, we sample 20 target images f and, for each target image, we sample 20 databases a^M. To evaluate predictive power of a model learned from a small $\widetilde{N}$-sized database $b^{\widetilde{N}}$ available for analysis, we use $\widetilde{N} = 41$.

B. Fast Nearest-Neighbor Search on Template Images

The proposed formulation is based on the independent variable being the deformed templates in the *entire* database. Multiatlas segmentation requires only a few most-similar templates (k in kNN) at each voxel in the image. To avoid a computationally-expensive nonlinear registration between the target image and *all* templates in the database, several strategies can be used. Some examples are as follows.

- The database can be organized using a hierarchical groupwise nonlinear-diffeomorphic registration scheme [31], [32] that produces several optimal mean templates for multiple classes of images within the database. A target can be mapped first to all the mean templates to determine the most similar classes and then the target can be registered to templates only within the similar classes.
- We can use fast approximate searches for similar templates relying on affine registration followed by spatial pyramid matching on coded geometry-capturing features, e.g., canny edges clustered and coded based on orientation and curvature [23]. The approximate search can be used to select a number of templates larger than the required $k^*(M)$ (e.g., twice or thrice $k^*(M)$), after which the selected templates can be nonlinearly registered to the target.

Both strategies result in very few nonlinear registrations. We perform nonlinear registration using a parallel algorithm for large deformation diffeomorphic metric mapping implemented on graphics processing units [33] available in AtlasWerks [34].

C. Size-Normalized Expected Segmentation Error $\mathcal{E}(M)$ and the Dice Similarity Coefficient

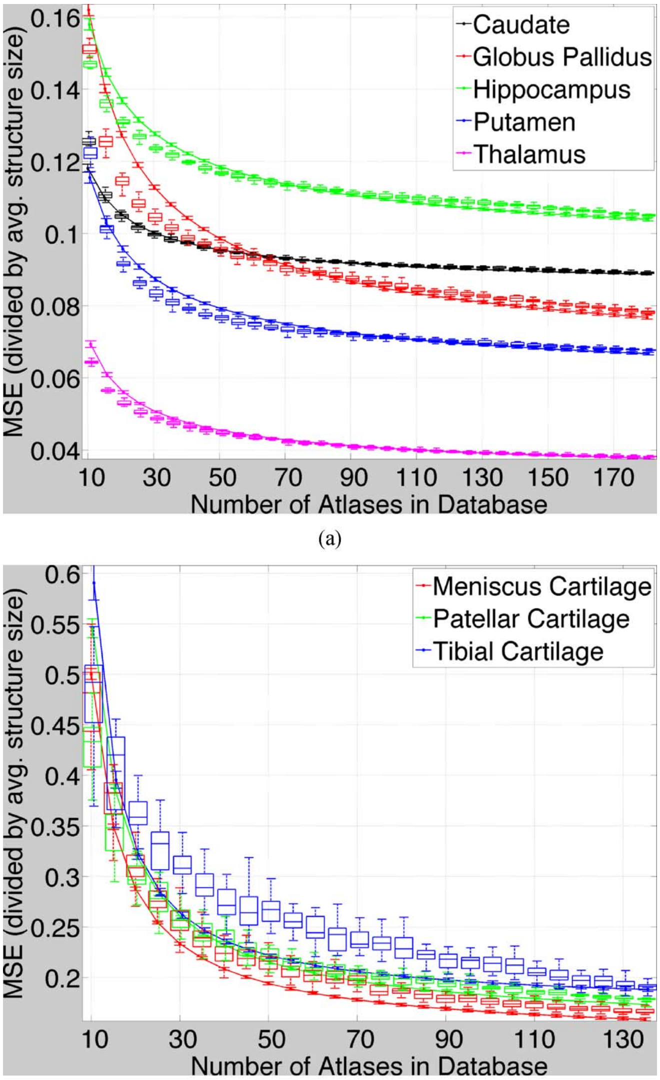
In this section, we motivate an alternative measure of segmentation performance, which *divides the MSE $\mathcal{E}(M)$ by the average true size of associated structure* in the database. This leads to a more meaningful interpretation as we find that, in practice, the size-normalized MSE values relate to the widely used Dice similarity coefficient (DSC) because both measure (dis)similarity between segmentations *relative to size*. Note that while DSC performs size normalization (via the denominator in the DSC formula) separately for every pair of segmentation images, the proposed approach rescales MSE values $\mathcal{E}(M), \forall M$, by dividing them by the same constant, i.e., the average true size of the structure in the database.

Theoretically, while DSC values lie within [0, 1], size-normalized MSE values lie within $[0, \eta]$, where η is twice the ratio of 1) the size of the largest structure in the database to 2) the average size of the structure in the database. Nevertheless, if the variability in the size of a structure is low, after the image is mapped to the common anatomical space, then the associated $\eta \rightarrow 2$ (this is generally observed in practice). Moreover, for binary segmentations s, s', the DSC is $2\sum_v (s_v s'_v) / \sum_v (s_v + s'_v)$ and the size-normalized MSE is $\sum_v (s_v + s'_v - 2 s_v s'_v)/\theta$, where θ is the average true size of the associated structure. When the size variability is low and most of the segmentation-image values are binary (generally observed in practice), then $\sum_v s_v \doteq \sum_v s'_v \doteq \theta$ and the size-normalized MSE becomes roughly $2(1 - \mathrm{DSC})$. We have found this formula to be a good approximation for the relationship between DSC and size-normalized MSE for real-world applications.

D. Validating the Parametric Model for Expected Segmentation Error $\mathcal{E}(M)$ as a Function of Database Size M

Fig. 1 shows 1) size-normalized MSE values, i.e., MSE values $\mathcal{E}(M, k)$ divided by the average true size of the structure in the database, and 2) the fitted parametric model, for various database sizes. Fig. 1 demonstrates that the parametric model fits the real-world data quite well.

Based on Fig. 1, the hippocampus is quite difficult to segment probably because of its elongated thin shape and small



(a)

(b)

Fig. 1. MSE (for fixed k) as a function of database size for (a) subcortical structures in brain MR images and (b) cartilages in knee MR images. Box-whisker plots show the *empirically computed* MSE values $\mathcal{E}(M_i, k = 8)$ after size normalization (see Section IV-C) for database sizes $M_i \in \{10, 15, 20, \ldots\}$ for the number of neighbors/templates in kNN fixed to eight independent of database size. Variance depicted by the whiskers comes from the bootstrap samples of databases $\{b_l^N\}_{l=1}^L$ available for analysis (we use $L = 20$). Solid lines and the associated error bars show the mean and standard deviation, respectively, of the *fitted* MSE values, using voxelwise parametric curve fitting with spatial regularization. The error bars come from the bootstrap samples of databases.

size. On the other hand, the thalamus gives the lowest MSEs probably due to its large size, despite the part of its boundary next to the gray matter being quite weak. The segmentation of cartilage structures in the knee is more challenging, as compared to the subcortical brain structures, probably because of the thin sheet-like shapes that are highly variable across the population represented in the database. Indeed, the knee MR images present a more challenging registration problem as compared to the brain dataset.

In Figs. 2–6 (b)–(d) and (f)–(h) show the mean and standard deviation, respectively, over Monte-Carlo bootstrap samples of databases $\{b_l^N\}_{l=1}^L$, of the parameter values $\{\alpha_v, \beta_v, d_v\}_{v=1}^V$ at each voxel v for the subcortical structures in brain MR images.

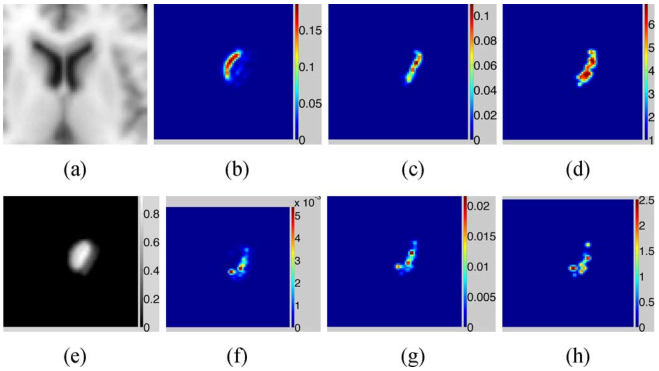


Fig. 2. Parameter values for multiatlas caudate segmentation from brain T1 MR images using large databases b_l^N. Images (a) and (e) show the average MR image and the average expert segmentation, respectively, in a common anatomical space. Images (b) and (f) show the mean and standard deviation, respectively, of α_v, at voxel v, generated from bootstrap samples of the databases $\{b_l^N\}_{l=1}^L$ employed for analysis. Similarly, (c) and (g) show the mean and standard deviation of β_v and (d) and (h) show the mean and standard deviation of d_v.

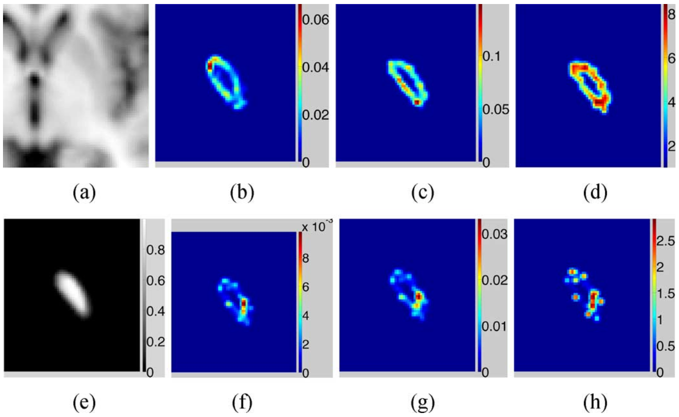


Fig. 3. Parameter values for multiatlas globus pallidus segmentation from brain T1 MR images using large databases b_l^N.

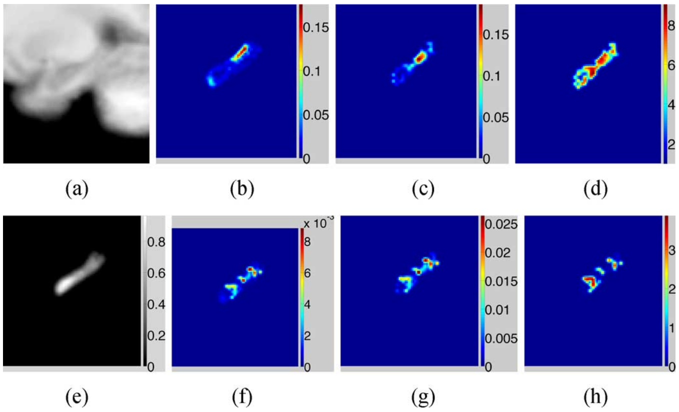


Fig. 4. Parameter values for multiatlas hippocampus segmentation from brain T1 MR images using large databases b_l^N.

We use $L = 20$. Similarly, in Figs. 7–9 (b)–(d) and (f)–(h) show the mean and standard deviation, respectively, of the parameter values for the cartilages in knee MR images. To estimate parameters, this paper uses $k_j \in \{5, 6, 7, 8, 9, 10\}$.

Values for α_v (inherent randomness) indicate the lowest possible MSE achievable with $k = 8$ and the chosen general-

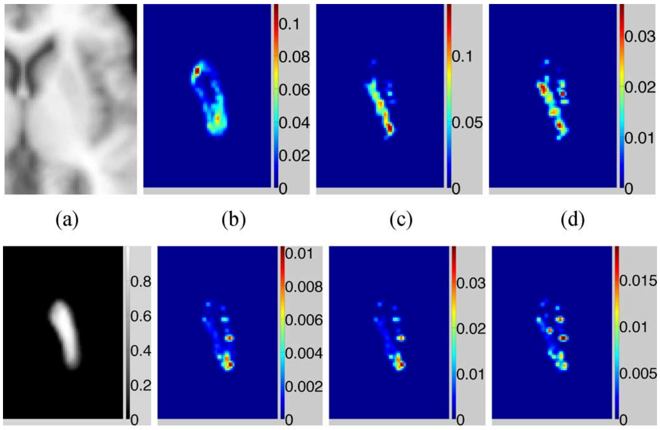


Fig. 5. Parameter values for multiatlas putamen segmentation from brain T1 MR images using large databases b_l^N.

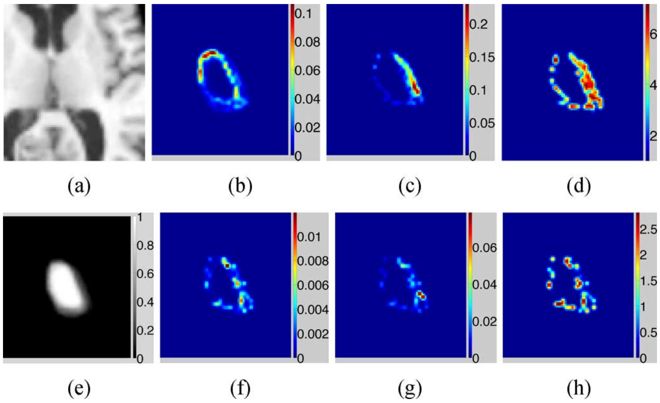


Fig. 6. Parameter values for multiatlas thalamus segmentation from brain T1 MR images using large databases b_l^N.

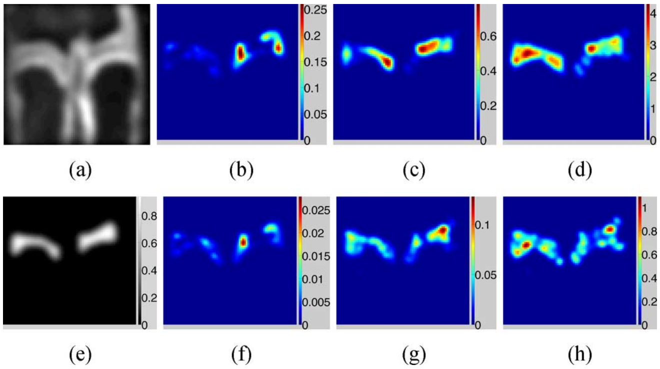


Fig. 7. Parameter values for multiatlas meniscus cartilage segmentation from knee T1 MR images using large databases b_l^N.

ized-kNN estimator. Values for β_v (regression complexity) and d_v (intrinsic dimension) are indicative of 1) the size of databases needed to achieve small MSEs, e.g., MSE closer to α_v, and 2) the amount of *benefit*, in terms of a decrease in MSE, obtained for the *cost* of an increase in database size. Such cost-benefit analyses are crucial for designing clinical support systems. Interestingly, the range of our estimates for d_v, for probabilistic segmentations, are roughly comparable to that found for several real-world datasets including fuzzy digit images and texture [24]–[28].

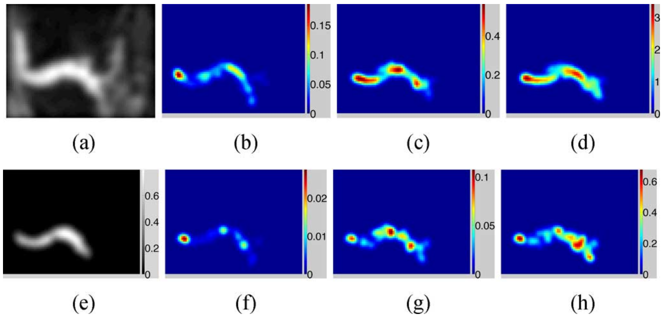


Fig. 8. Parameter values for multiatlas patellar cartilage segmentation from knee T1 MR images using large databases b_l^N.

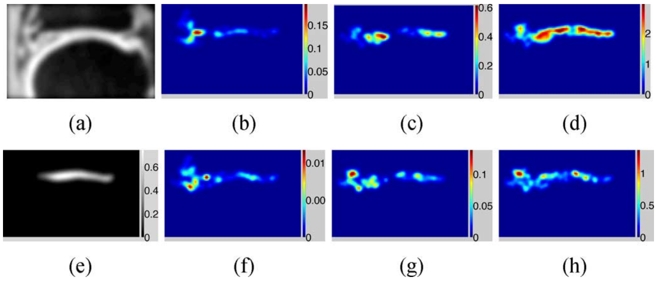
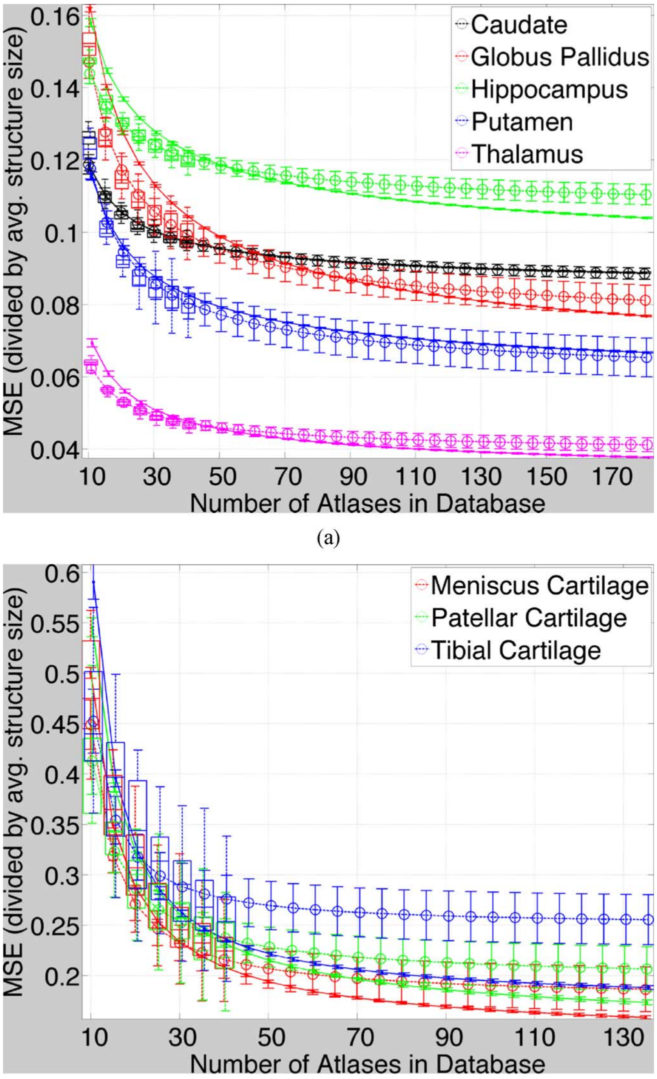


Fig. 9. Parameter values for multiatlas tibial cartilage segmentation from knee T1 MR images using large databases b_l^N.

For voxels v well inside or well outside the structures, the values of α_v and β_v are close to zero, which implies that the MSE values are also close to zero. As expected, this indicates the ease of segmentation for such voxels. Voxels where the segmentation is the most difficult (high values for α_v, β_v, d_v) lie near the boundaries of the structures, especially near those boundaries that are weakly visible in the MR image *and* where the boundary locations cannot be well predicted based on the locations of nearby landmarks that are clearly visible in the images. Prominent examples of such voxels are the tail of the hippocampus and some parts of the interfaces of the caudate, globus pallidus, putamen, and thalamus with the white matter. Interestingly, while the weakly visible posterior boundaries of putamen (where the shape tapers off) are difficult to delineate, the weakly visible anterior boundaries of the putamen are much easier to delineate. Similarly, the weakly visible hippocampus head, where the hippocampus touches the amygdala, is easier to predict, based on nearby clearly visible landmarks, than the hippocampus tail.

E. Validating the Predictive Model of Expected Segmentation Error $\mathcal{E}(M)$ for Large Database Sizes M

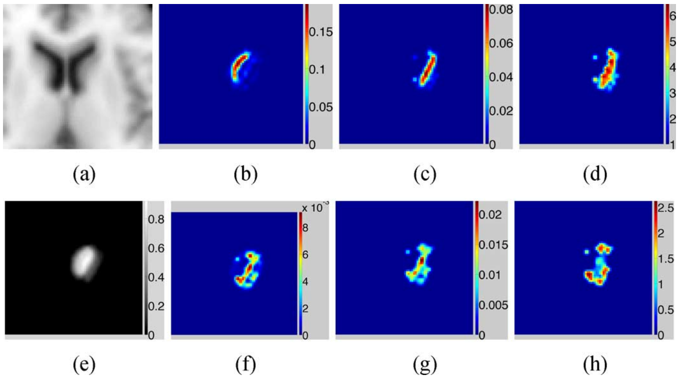
Fig. 10 shows the results of experiments using Monte-Carlo bootstrap samples of smaller databases $\{b_l^{\widetilde{N}}\}_{l=1}^{L}$; we use $\widetilde{N} = 41$, $L = 20$. Fig. 10 shows that the predicted MSE for large databases ($M \gg \widetilde{N}$) (predicted using small databases) is quite close (typically within one standard deviation) to the fitted MSE values computed using large databases b_l^N. This validates the proposed predictive model. Thus, small databases,



(a)

(b)

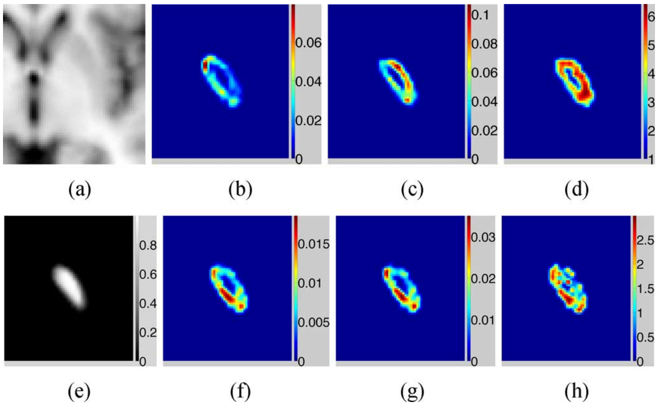
Fig. 10. Predicting MSE (for fixed k) as a function of database size for (a) subcortical Structures in brain MR images and (b) cartilages in knee MR Images. Box-whisker plots show the *empirically computed* MSE values $\mathcal{E}(M_i, k=8)$ after size normalization (see Section IV-C) for $M_i \in \{10, 15, 20, \ldots\}$ for the number of neighbors/templates in kNN fixed to eight independent of database size. Variance depicted by the whiskers comes from the bootstrap samples of small databases $\{b_l^{\widetilde{N}}\}_{l=1}^{L}$ available for analysis (we use $\widetilde{N}=41$, $L=20$). Solid lines and associated error bars are the same as those in Fig. 1 obtained using the larger databases b_l^N. Dashed lines and associated error bars show the mean and standard deviation, respectively, of the *fitted* MSE values using small databases $b_l^{\widetilde{N}}$. Error bars come from the bootstrap samples of databases.



(a) (b) (c) (d)

(e) (f) (g) (h)

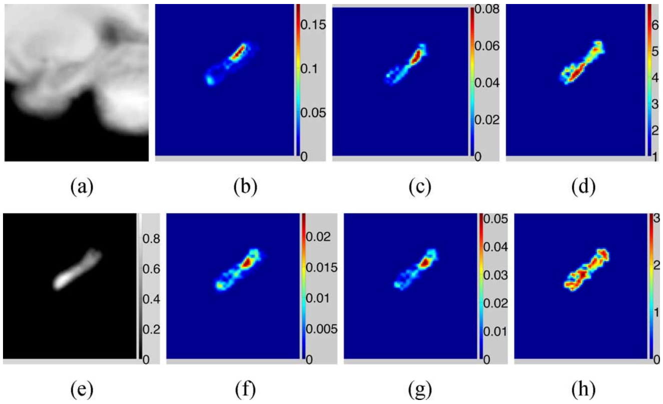
Fig. 11. Parameter values for multiatlas caudate segmentation from brain T1 MR images using small databases $b_l^{\widetilde{N}}$. Images (a) and (e) show the average MR image and the average expert segmentation, respectively, in a common anatomical space. These are the same as those shown in the previous section, in Fig. 2. Images (b) and (f) show the mean and standard deviation, respectively, of α_v, at voxel v, generated from bootstrap samples of the small databases $\{b_l^{\widetilde{N}}\}_{l=1}^{L}$ employed for parameter estimation (we use $\widetilde{N}=41$, $L=20$). Similarly, (c) and (g) show the mean and standard deviation of β_v and (d) and (h) show the mean and standard deviation of d_v.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 12. Parameter values for multiatlas globus pallidus segmentation from brain T1 MR images using small databases $b_l^{\widetilde{N}}$.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 13. Parameter values for multiatlas hippocampus segmentation from brain T1 MR images using small databases $b_l^{\widetilde{N}}$.

which require fewer expert segmentations and less time and effort to construct, can be used to predict the much larger database sizes required to achieve a specified maximum tolerable error in segmentation. Such cost-benefit analysis is crucial for designing and deploying multiatlas segmentation systems, potentially comprising a few thousand atlases.

Overall, the quality of prediction is a little better for subcortical structures in the brain MR images as compared to the cartilages in the knee MR images. This is consistent with the findings in the previous section.

In Figs. 11–15 (b)–(d) and (f)–(h) show the mean and standard deviation, respectively, over Monte-Carlo bootstrap sampling of databases b_l^N, of the parameter values $\{\alpha_v, \beta_v, d_v\}_{v=1}^{V}$ at each voxel v for the subcortical structures in brain MR images. Similarly, in Figs. 16–18 (b)–(d) and (f)–(h) show the mean and standard deviation, respectively, of the parameter

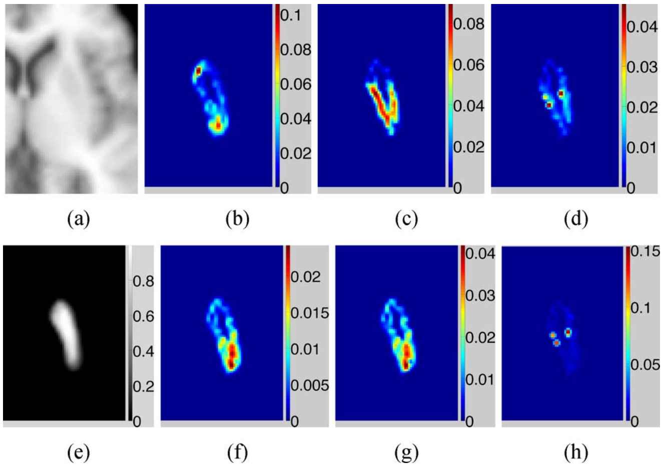


Fig. 14. Parameter values for multiatlas putamen segmentation from brain T1 MR images using small databases $\widetilde{b_l^N}$.

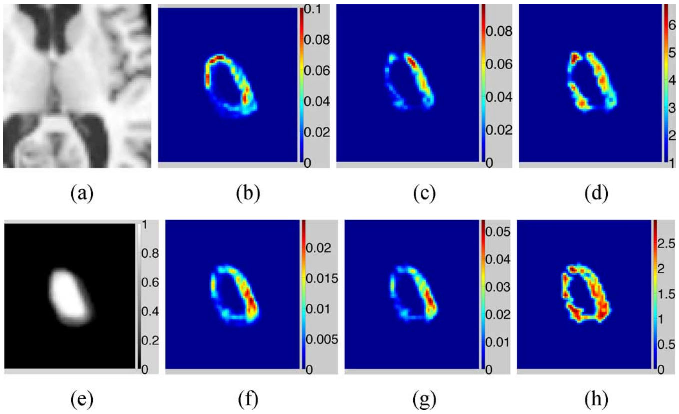


Fig. 15. Parameter values for multiatlas thalamus segmentation from brain T1 MR images using small databases $\widetilde{b_l^N}$.

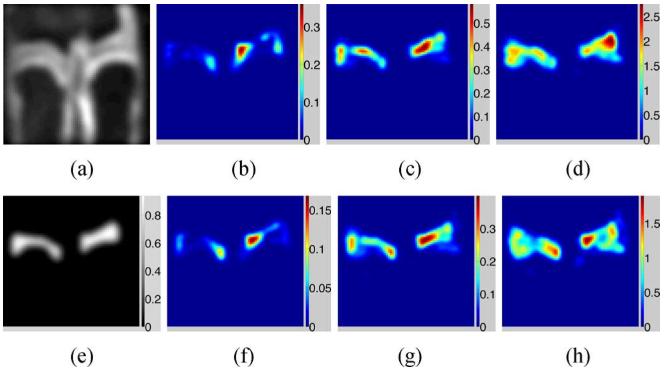


Fig. 16. Parameter values for multiatlas meniscus cartilage segmentation from knee T1 MR images using small databases $\widetilde{b_l^N}$.

values for the cartilages in knee MR images. These parameter values obtained from small databases are a good approximation to those obtained using large databases in Section IV-D, To estimate parameters, this paper uses $k_j \in \{5, 6, 7, 8, 9, 10\}$.

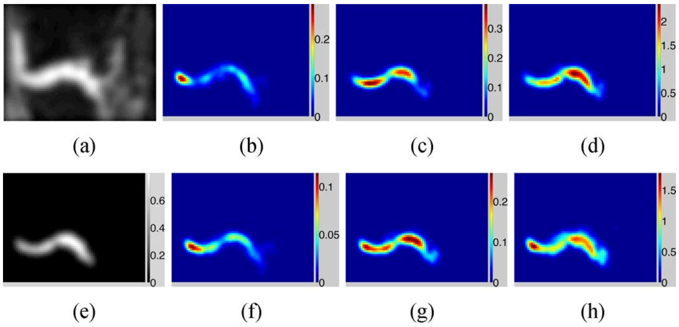


Fig. 17. Parameter values for multiatlas patellar cartilage segmentation from knee T1 MR images using small databases $\widetilde{b_l^N}$.

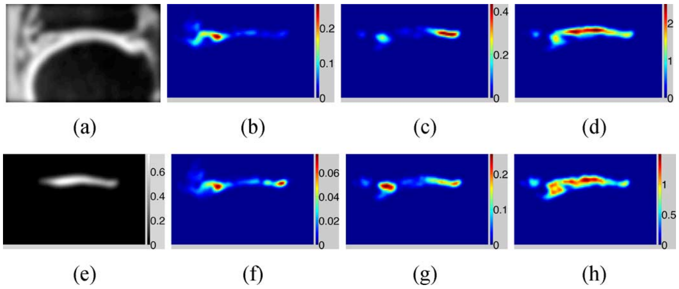
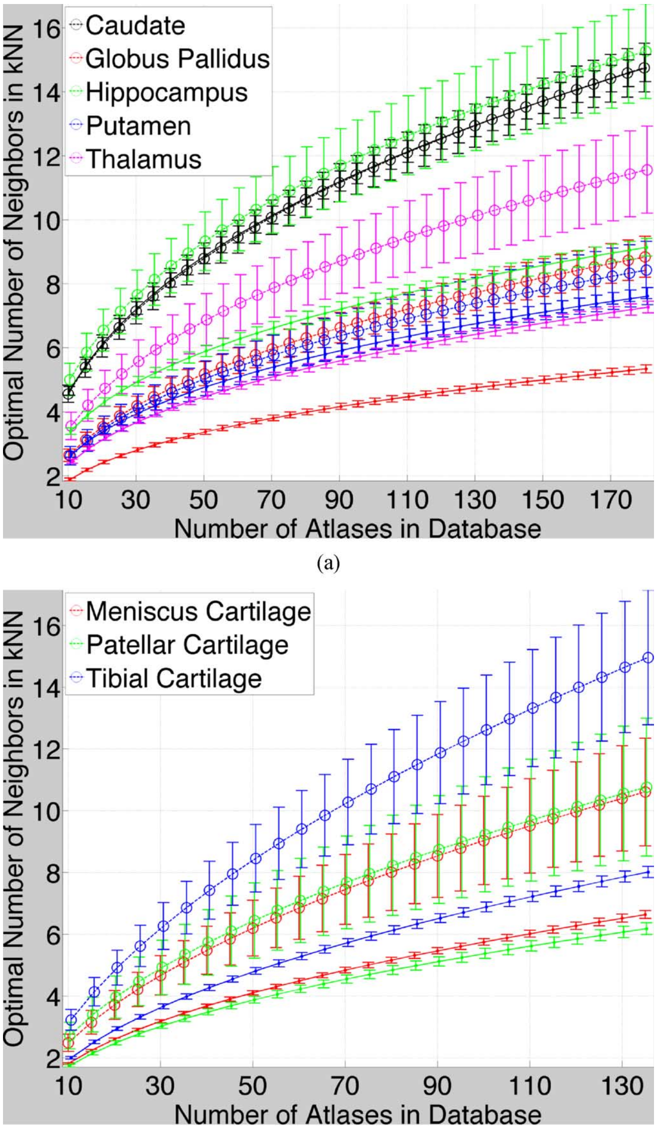


Fig. 18. Parameter values for multiatlas tibial cartilage segmentation from knee T1 MR images using small databases $\widetilde{b_l^N}$.

F. Validating Regression-Estimator Optimization (Optimal Number of Nearest Templates $k^*(M)$) and Resulting Prediction of Expected Segmentation Error $\mathcal{E}(M)$

Fig. 19 shows the optimal number of nearest neighbors/templates $k^*(M)$ computed from the parameter fits. As $M \to \infty$, we see that $k^*(M) \to \infty$ and $k^*(M)/M \to 0$, thereby ensuring convergence of the kNN estimators at each voxel to the true conditional-expectation regression function. For some brain structures, for large M, $k^*(M)$ predicted using small databases $\widetilde{b_l^N}$ is within 10% of the $k^*(M)$ computed using large databases b_l^N. Even though the optimal values $k^*(M)$ for many structures computed from small and large databases are *not* that close in terms of the standard deviations, the absolute differences between them are only of the order of a few number of templates. Nevertheless, the MSE prediction relying on the optimized $k^*(M)$ is largely unaffected. Indeed, Fig. 20 shows that the predictions of MSE values, for large M, via parameter estimation and optimization of $k^*(M)$ using much smaller databases, match the MSE values computed using large databases quite well.

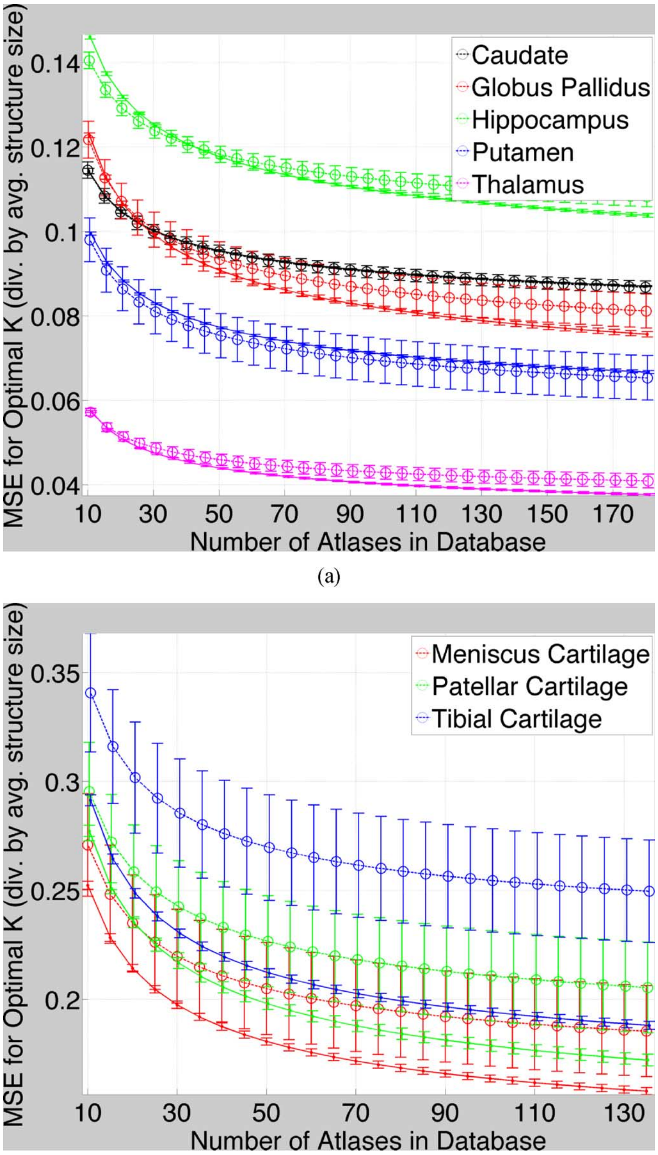
More importantly, the MSE values with optimal $k^*(M)$ in Fig. 20 are lower than those with a fixed k in Fig. 10, for small values of the database size M. Thus, the proposed algorithm for optimally estimating $k^*(M)$ and using that for multiatlas segmentation performs better than the strategy where k is fixed independent of the database size M.



(a)

(b)

Fig. 19. Predicting optimal $k^*(M)$ for large database sizes using small databases for (a) subcortical structures in brain MR images and (b) cartilages in knee MR images. Solid lines and associated error bars, show the mean and standard deviation, respectively, of the optimal number of nearest neighbors/templates $k^*(M)$ computed using large bootstrap-sampled databases b_l^N. Similarly, dotted lines and associated error bars, show the same quantities computed using small bootstrap-sampled databases $b_l^{\widetilde{N}}$. Error bars come from the bootstrap samples of databases.



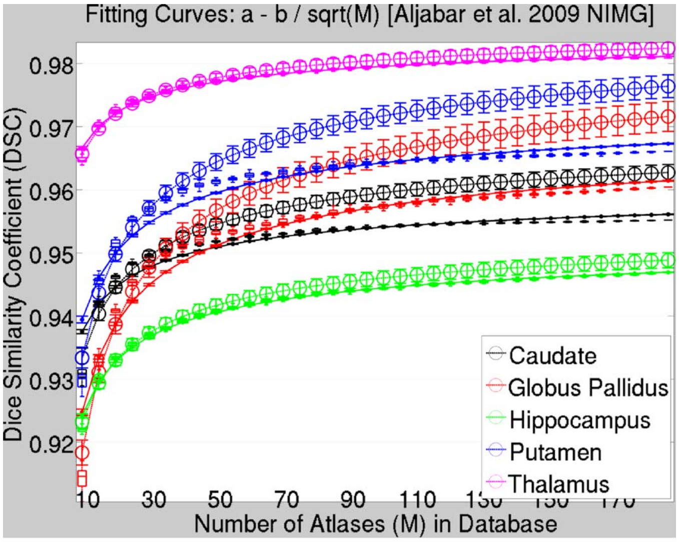
(a)

(b)

Fig. 20. Predicting MSE, using optimal $k^*(M)$, for large database sizes using small databases for (a) subcortical structures in brain MR images and (b) cartilages in knee MR images. Solid lines and associated error bars, show the mean and standard deviation, respectively, of the MSE (after the size normalization described in Section IV-C) computed using the optimal number of nearest neighbors/templates $k^*(M)$ for large bootstrap-sampled databases b_l^N. Similarly, dotted lines, and associated error bars, show the same quantities computed using small bootstrap-sampled databases $b_l^{\widetilde{N}}$. Error bars come from the bootstrap samples of databases.

G. Comparing Segmentation-Performance Prediction With an Earlier Model [3]

This section presents results using an earlier pioneering approach in the literature [3], [14], [15], which comes closest to our proposed approach, for modeling multiatlas-segmentation performance parametrically. The approach in [3] is restricted to a global analysis, by being based on the DSC measure, and models the DSC, for a given database size M, as the curve $a - b/\sqrt{M}$, where a and b are model parameters. This parametric model was motivated by the need to quantify suboptimality in DSC measures resulting from random errors (via parameter b) and systematic errors (via parameter a), where the errors might stem from, e.g., misregistration or atlas inconsistencies. Unlike our proposed approach, the parametric model was perhaps *not* intended to be used for measuring segmentation performance as a function of database size and may *not* be perfectly applicable when the atlases are selected in a data driven manner [3]. Nevertheless, the model appears to be more general, having wider utility, and does suggest that for large database sizes M, the effect of the random errors is nullified. In this spirit, we choose to compare our proposed approach with the parametric model in [3] for predicting multiatlas-segmentation performance for large databases, where the predictive



(a)

Fig. 21. Comparing segmentation-performance prediction with an earlier model [3]: predicting dice similarity for large database sizes using small databases for (a) subcortical structures in brain MR images. Box-whisker plots show the *empirically computed* DSC values $\mathrm{DSC}(M_i)$ for $M_i \in \{10, 15, 20, \ldots\}$ for the number of templates fixed to 8, independent of database size M_i. Variance depicted by the whiskers comes from the bootstrap samples of small databases $\{b_l^{\widetilde{N}}\}_{l=1}^{L}$ available for analysis (we use $L = 20$). *Solid* lines and the associated error bars show the mean and standard deviation, respectively, of the *fitted* DSC values, using the *larger* databases b_l^N; the error bars come from the bootstrap samples of databases. *Dashed* lines and associated error bars show the mean and standard deviation, respectively, of the *fitted* DSC values using *small* databases $b_l^{\widetilde{N}}$; the error bars come from the bootstrap samples of databases.

model is built using a small database. We perform a rigorous fair Monte-Carlo analysis for comparing segmentation-performance predictability with our approach (Fig. 10).

Fitting the model in [3] leads to a weighted linear least-squares optimization problem (unlike the nonlinear problem for our approach in Section III-D), where the optimal estimates for the parameters a, b are obtained in closed form by minimizing the sum of (weighted) squared deviations between the model-dictated DSC curve and the observed DSC values; where the weights are the inverse of the computed variances associated with the set of squared deviations, via bootstrap sampling (analogous to our approach).

The results of the approach in [3] in Fig. 21 clearly show large errors in predicting the DSC for three of the five subcortical structures, i.e., caudate, globus pallidus, and putamen. For instance, for [3] in Fig. 21, the mean and median DSC values (for fitted solid lines and empirical box plots, respectively) using large-sized databases deviate from the predicted mean DSC by more than 3–4 standard deviations of the predicted DSC. In contrast, for the proposed approach in Fig. 10, the mean and median segmentation errors (for fitted solid lines and empirical box plots, respectively) using large-sized databases deviate from the predicted segmentation error by at-most 1 standard deviation (mostly far less than that) of the predicted segmentation error. Thus, the proposed approach has significantly better predictive power than that in [3] for the subcortical structures in this large brain MR image database.

It is clear that the major factor for the significantly improved performance of the proposed method stems from its ability to optimize the (intrinsic) dimension parameter d_v; *not* just for each structure overall, but for every voxel within each structure. On the other hand, the approach in [3], to force an analogy, fixes d_v to 8 for all voxels v for every structure. Thus, the predictive model of [3] incidentally works well for some structures (hippocampus and thalamus), but fails for all others. In contrast, our approach presents a principled sophisticated model for predicting segmentation performance, which offers the flexibility of tuning the dimension parameter for each voxel for each anatomical structure, thereby yielding a significantly improved prediction for segmentation performance. Furthermore, the proposed approach offers other significant benefits to improve prediction, as discussed in Section II, including the ability 1) to model segmentation error as a function of the size k of the subset selected for averaging and 2) to optimize the subset size as a function of the database size.

V. Discussion and Conclusion

This paper establishes a brand-new principled theoretical framework for the modeling and analysis of multiatlas segmentation relying on local generalized-kNN nonparametric regression and Mercer-kernel-based distances between images. It shows how the proposed framework can be utilized to measure the difficulty of a specific multiatlas segmentation problem in terms of the convergence behavior of expected segmentation error as a function of database size. It captures these properties using parameters that are fundamental to the underlying generalized-kNN regression model for multiatlas segmentation. These parameters capture the natural variance of the PDFs of the segmentation conditioned on the templates and the bias and variance of the kNN estimator. It also uses these parameters to optimize the regression estimator. It shows the validity of the model using rigorous Monte-Carlo validation on two large clinical databases. This paper shows how the proposed framework coupled with a small set of atlases (requiring few expert segmentations) can be utilized to predict the much-larger database sizes (“cost”) required to achieve a specified maximum tolerable error (“benefit”) in segmentation. Such “cost-benefit” analysis is crucial for designing and deploying multiatlas segmentation systems comprising, potentially, several thousands of atlases.

Apart from presenting a novel theoretical framework for analysis of multiatlas segmentation, this paper also offers novelty in the context of label-fusion algorithms through the proposed method for estimating the optimal subset of atlases, with subset size $k^*(M)$ being a function of database size M, and using this optimal subset for segmentation.

The quality of prediction of expected segmentation error depends on the size $\widetilde{N}$ of the small database and the factors determining the *specific segmentation problem*. Our results indicate that the prediction quality for the brain structures may be unaffected with database sizes somewhat smaller than what we used ($\widetilde{N} = 41$), but the prediction for the knee structures may become substantially worse with smaller $\widetilde{N}$.

The proposed framework allows the template images to be registered to the common anatomical space, or the target image space, as long as the warps are diffeomorphic. The assumption of the images warped to a common anatomical space is mainly theoretical in nature and is *not* practically restrictive because the warps are assumed to be diffeomorphic (i.e., smooth and *invertible*). Thus, under the assumption of diffeomorphic registrations, (theoretically) registering to the common anatomical space and computing errors in that space is equivalent to registering to the target space and evaluating segmentation errors in the target space. In practice, label fusion in the target space might produce better segmentations.

The current theoretical framework relies on diffeomorphisms. Nevertheless, this is *not* a major concern because of the increasing popularity and practical utility of diffeomorphic registration. How well would the proposed analysis behave (e.g., how robust would it be) if the strict diffeomorphic-registration assumption is lifted would be interesting to study in the future. We believe that, because the popular methods for nonlinear nondiffeomorphic registration yield deformations that are reasonably close to being diffeomorphic, 1) the results of such nondiffeomorphic registration methods are quite similar to those using nonlinear diffeomorphic registration and 2) the results of multiatlas segmentation and the predictive analysis in the two scenarios would also remain quite similar.

Taking a broader perspective, the proposed framework can be employed for modeling and analysis of approaches where the segmentation-image dependent variable, underlying the regression, is replaced by other kinds of clinical data, e.g., clinically relevant test scores.

Appendix

Theorem 1: (Local Normalized Cross Correlation is a Mercer kernel) $\mathcal{K}_v(\cdot,\cdot)$*, defined in (15), is a Mercer kernel.*

Proof: Consider any finite set of vectorized image observations $\{e^m \in \mathbb{R}^V\}_{m=1}^M$. Consider this set represented as a $V \times M$ matrix E such that the mth column of E is e^m.

For the function $\mathcal{K}_v(\cdot,\cdot)$ at voxel v, consider a patch around v comprising $1 \leq p \leq V$ voxels. Without loss of generality, consider that the vectorized images e^m are permuted such that the p intensities in the patch are the first p components (in some fixed order) of e^m or the first p rows in E.

Consider a $p \times V$ diagonal matrix A_v, with all elements along the diagonal being 1. Then, $A_v e$ crops the vectorized image e and gives the intensities in the patch around voxel v.

Consider a $p \times p$ matrix B_v with all elements having value $1/p$. Then $(\mathbf{1} - B_v)A_v e$ gives the mean-subtracted patch intensities. Let $C_v := (\mathbf{1} - B_v)A_v$.

Consider any vector of real numbers $\Lambda \in \mathbb{R}^M$ where the mth component of Λ is denoted by Λ_m. Then,

$$\begin{aligned}\sum_{m=1}^{M}\sum_{m'=1}^{M}\Lambda_m\Lambda_{m'}\widetilde{\mathcal{K}}_v(e^m, e^{m'}) &= \sum_{m,m'}\Lambda_m\Lambda_{m'}\langle C_v e^m, C_v e^{m'}\rangle \\ &= \langle C_v E\Lambda, C_v E\Lambda\rangle \\ &= \|C_v E\Lambda\|_2^2 \geq 0 \qquad (23)\end{aligned}$$

for any chosen Λ and set of vectorized images $\{e^m\}_{m=1}^M$. The function $\widetilde{\mathcal{K}}_v(\cdot,\cdot)$ is also continuous and symmetric. Therefore, the local cross correlation $\widetilde{\mathcal{K}}_v(\cdot,\cdot)$, defined in (16), is a Mercer kernel [35]. The function of interest $\mathcal{K}_v(\cdot,\cdot)$, defined in (15), is the normalized version of the Mercer kernel $\widetilde{\mathcal{K}}_v(\cdot,\cdot)$. Therefore, $\mathcal{K}_v(\cdot,\cdot)$ is a Mercer kernel [35]. Q.E.D. ■

Acknowledgment

This manuscript was prepared using an OAI public use data set and does not necessarily reflect the opinions or views of the OAI investigators, the NIH, or the private funding partners. The authors would like to thank P. Zhu for help with data preprocessing and the anonymous reviewers for providing insightful feedback.

References

[1] M. Cabezas, A. Oliver, Z. Llado, J. Freixenet, and M. Cuadra, "A review of atlas-based segmentation for magnetic resonance brain images," *Comput. Methods Programs Biomed.*, vol. 104, no. 3, pp. 158–177, 2011.

[2] D. Pham, C. Xu, and J. Prince, "Current methods in medical image segmentation," *Annu. Rev. Biomed. Eng.*, vol. 2, pp. 315–337, 2000.

[3] P. Aljabar, R. Heckemann, A. Hammers, J. Hajnal, and D. Rueckert, "Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy," *NeuroImage*, vol. 46, no. 3, pp. 726–738, 2009.

[4] M. Depa, M. R. Sabuncu, G. Holmvang, R. Nezafat, E. J. Schmidt, and P. Golland, "Robust atlas-based segmentation of highly variable anatomy: Left atrium segmentation," in *MICCAI Workshop Stat. Atlases Comp. Models Heart*, 2010, pp. 1–8.

[5] J. Lotjonen, R. Wolz, J. Koikkalainen, L. Thurfjell, G. Waldemar, H. Soininen, and D. Rueckert, "Fast and robust multi-atlas segmentation of brain magnetic resonance images," *NeuroImage*, vol. 49, no. 3, pp. 2352–2365, 2010.

[6] X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de-Solorzano, "Combination strategies in multi-atlas image segmentation: Application to brain MR data," *IEEE Trans. Med. Imag.*, vol. 28, no. 8, pp. 1266–1277, Aug. 2009.

[7] A. Duc, M. Modat, K. Leung, M. Cardoso, J. Barnes, T. Kadir, and S. Ourselin, "Using manifold learning for atlas selection in multi-atlas segmentation," *PLOS ONE*, vol. 8, no. 8, p. e70059, 2013.

[8] I. Isgum, M. Staring, A. Rutten, M. Prokop, M. Viergever, and B. Ginneken, "Multi-atlas-based segmentation with local decision fusion–Application to cardiac and aortic segmentation in CT scans," *IEEE Trans. Med. Imag.*, vol. 28, no. 7, pp. 1000–1010, Jul. 2009.

[9] H. Jia, P.-T. Yap, and D. Shen, "Iterative multi-atlas-based multi-image segmentation with tree-based registration," *NeuroImage*, vol. 59, no. 1, pp. 422–430, 2012.

[10] M. Sabuncu, B. Yeo, K. van Leemput, B. Fischl, and P. Golland, "A generative model for image segmentation based on label fusion," *IEEE Trans. Med. Imag.*, vol. 29, no. 10, pp. 1714–1729, Oct. 2010.

[11] H. Wang, J. Suh, S. Das, J. Pluta, C. Craige, and P. Yushkevich, "Multi-atlas segmentation with joint label fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 611–623, Mar. 2012.

[12] H. Wang, J. W. Suh, J. Pluta, M. Altinay, and P. Yushkevich, "Optimal weights for multi-atlas label fusion," in *Proc. Int. Conf. Info. Med. Imag.*, 2011, pp. 73–84.

[13] S. P. Awate, P. Zhu, and R. T. Whitaker, "How many templates does it take for a good segmentation?: Error analysis in multiatlas segmentation as a function of database size," in *Proc. Int. Workshop Multimodal Brain Image Analys. Int. Conf. Med. Image Comput. Comp. Assist. Interv.*, 2012, Lecture Notes Comput. Sci., pp. 103–114.

[14] R. Heckemann, J. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Automatic anatomical brain MRI segmentation combining label propagation and decision fusion," *NeuroImage*, vol. 33, no. 1, pp. 115–126, 2006.

[15] R. Heckemann, J. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers, "Multiclassifier fusion in human brain MR segmentation: Modelling convergence," in *Med. Image. Comput. Comp. Assist. Interv.*, 2006, pp. 815–822.

[16] M. Sabuncu, S. K. Balci, M. E. Shenton, and P. Golland, "Image-driven population analysis through mixture modeling," *IEEE Trans. Med. Imag.*, vol. 28, no. 9, pp. 1473–1487, Sep. 2009.

[17] S. Warfield, K. Zou, and W. Wells, "Validation of image segmentation by estimating rater bias and variance," *Phil. Trans. R. Soc.*, vol. 366, no. 1874, pp. 2361–2375, 2008.

[18] O. Commonwick and S. Warfield, "Estimation of inferential uncertainty in assessing expert segmentation performance from STAPLE," *IEEE Trans. Med. Imag.*, vol. 29, no. 3, pp. 771–780, Mar. 2010.

[19] W. Hardle, *Applied Nonparametric Regression*. Cambridge, U.K.: Cambridge Univ. Press, 1990.

[20] Y. P. Mack, "Local properties of k-NN regression estimates," *SIAM J. Alg. Disc. Meth.*, vol. 2, no. 3, pp. 311–323, 1981.

[21] K. Grauman and T. Darrell, "The pyramid match kernel: Efficient learning with sets of features," *J. Mach. Learn. Res.*, vol. 8, pp. 725–760, 2007.

[22] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2006, pp. 2169–2178.

[23] P. Zhu, S. P. Awate, S. Gerber, and R. T. Whitaker, "Fast shape-based nearest-neighbor search for brain MRIs using hierarchical feature matching," in *Proc. Med. Image Comput. Comp. Assist. Interv.*, 2011, vol. 2, pp. 484–491.

[24] K. Carter, R. Raich, and A. Hero, "On local intrinsic dimension estimation and its applications," *IEEE Trans. Signal Process.*, vol. 58, no. 2, pp. 650–663, Feb. 2010.

[25] D. de Ridder, O. Kuoropteva, O. Okun, M. Pietikainen, and R. Duin, "Supervised locally linear embedding," in *Proc. Int. Conf. Artif. Neural Netw.*, 2003, pp. 333–341.

[26] M. Felsberg, S. Kalkan, and N. Krueger, "Continuous dimensionality characterization of image structures," *Image Vis. Comput.*, vol. 27, no. 6, pp. 628–636, 2009.

[27] M. Hein and J.-Y. Audibert, "Intrinsic dimensionality estimation of submanifolds in $\mathbb{R}^d$," in *Proc. Int. Conf. Mach. Learn.*, 2005, pp. 289–296.

[28] M. Raginsky and S. Lazebnik, "Estimation of intrinsic dimensionality using high-rate vector quantization," in *Proc. Adv. Neural Informat. Process. Syst.*, 2005, pp. 1–8.

[29] B. Efron, "Bootstrap methods: Another look at the jackknife," *Ann. Stat.*, vol. 7, no. 1, pp. 1–26, 1979.

[30] P. C. Hansen, P. Johnston, Ed., "The L-curve and its use in the numerical treatment of inverse problems," in *Computat. Inverse Problems Electrocardiol.*, 2001, pp. 119–142.

[31] M. Sabuncu, M. Shenton, and P. Golland, "Joint registration and clustering of images," in *Med. Image. Comput. Comp. Assist. Intervent.*, 2007, vol. 10, pp. 47–54.

[32] Q. Wang, L. Chen, P.-T. Yap, G. Wu, and D. Shen, "Groupwise registration based on hierarchical image clustering and atlas synthesis," *Human Brain Mapp.*, vol. 31, no. 8, pp. 1128–1140, 2010.

[33] L. Ha, J. Kruger, T. Fletcher, S. Joshi, and C. Silva, "Fast parallel unbiased diffeomorphic atlas construction on multi-graphics processing units," in *Eur. Symp. Parallel Graph. Vis.*, 2009, pp. 65–72.

[34] AtlasWerks: An Open-Source (BSD License) Software Package for Medical Image Atlas Generation SCI Inst. Sci. Comput. Imag. Inst., 2013 [Online]. Available: http://www.sci.utah.edu/software/atlaswerks.html

[35] B. Scholkopf and A. Smola, *Learning with Kernels*. Cambridge, MA: MIT Press, 2002.