To Appear: Information Processing in Medical Imaging (IPMI 2017)

From Label Maps to Generative Shape Models: A Variational Bayesian Learning Approach

Shireen Y. Elhabian and Ross T. Whitaker

School of Computing, University of Utah, Salt Lake City, UT, USA

Abstract. Shape models of grid-structured representations, aka *label maps*, are a useful tool for a variety of medical imaging tasks, including segmentation and shape analysis. In this paper, we propose a Bayesian treatment of a latent variable model for learning generative shape models, called *ShapeOdds*, that relies on direct probabilistic formulation with a variational approach for deterministic model learning. Spatial coherency and sparsity priors are also incorporated to lend stability to the optimization problem, thereby regularizing the solution space while avoiding overfitting in this high-dimensional, low-sample-size scenario. We deploy a type-II maximum likelihood estimate of the model hyperparameters to avoid grid searches. Further, we propose a *mixture* variant of ShapeOdds for capturing nonlinear shape variations in a way that balances the model expressiveness with the efficiency of learning and inference. Experiments show that the proposed model outperforms state-of-the-art representations on real datasets w.r.t. generalization to unseen samples.

Keywords: Statistical shape models, variational inference, mixture modeling, LogOdds, Generalized Expectation-Maximization

1 Introduction

Statistical shape modeling has gained wide acceptance in medical imaging applications for representing variability within populations of similar or related anatomical shapes. By providing a compact parameterization of the underlying anatomical variation, shape models have important applications in a variety of medical imaging tasks, such as image segmentation (*e.g.,* [1]), hypothesis testing (*e.g.,* [2]), anatomical reconstruction (*e.g.,* [3]), and pathology detection (*e.g.,* [4]). In image segmentation, for instance, sources of uncertainty in generative image models often include appearance and shape. In the presence of misleading appearance information, one typically incorporates a prior distribution for the expected object shape to bias the solution toward a shape-class of interest (*e.g.,* [5,6]). Such a prior helps overcome a range of challenges with the image data, including low signal-to-noise ratio and ill-defined boundaries.

The problem of modeling shape variation is often posed as estimating a probability distribution from a set of training samples drawn from an unknown distribution, treating individual *i.i.d.* samples as data points in a high-dimensional *shape space*. This paper focuses on learning generative shape models from grid-structured data that define a binary function over a discrete image domain,

aka *label maps*. Learning shape models in the space of label maps is challenging because the space is a *unit hyper-cube* and lacks a vector space structure. Thus, the binary variables entail *non-Gaussian data likelihood*, which leads to intractable marginals and posteriors. The combination of *high-dimensional shape space* and limited training samples presents a risk of overfitting. *Hyperparameters* associated with model complexity often result in computationally expensive grid searches. Furthermore, nonlinear shape variation is often rendered as *multimodal probability distributions*, which confound the many approaches that rely on Gaussian assumptions (*e.g.,* [5, 7, 8]). Nonparametric density estimates (*e.g.,* [6,9,10]) are promising, but have a ravenous appetite for training samples, posing serious challenges for the robustness and generalization, with an inherent tradeoff between the estimate bias and variance [11].

To handle non-Gaussian data likelihoods associated with label maps, most existing approaches have resorted to learning shape distributions *indirectly*, by representing shapes as level sets, and working in the space of continuous implicit functions, such as blurred label maps (*e.g.,* [7,12]) or signed distance maps (SDMs) (*e.g.,* [7,13]). These ad hoc embeddings lack a statistical interpretation w.r.t. the input label maps, which leads to suboptimal generative models and limits their ability to generalize to unseen data. On the other hand, learning shape variability *directly* on the label map space (*e.g.,* [14]) or a probabilistic label space (*e.g.,* [8]) using, for instance, principal component analysis (PCA) has no underlying generative model, and therefore does not benefit from associated tools for statistical estimation. For instance, sampling along PCA modes often results in invalid shapes that do not belong to the unit hyper-cube shape space, requiring projections (*e.g.,* clamping or thresholding) that typically discard associated estimates of uncertainty. Moreover, non-Bayesian dimensionality reduction techniques (*e.g.,* linear [5, 7, 8] and nonlinear [15]) rely on the maximum-likelihood estimation of the principal subspace, and thus ignore uncertainties associated with the estimated low-dimensional representation and are prone to overfitting in high-dimensional spaces with limited training samples [16]—a typical situation in medical applications with small—and large-scale shape variability. Recently, stochastic neural nets, in particular the shape Boltzmann machine (ShapeBM) [17], have been adopted to learn distributions over binary inputs without relying on any intermediate implicit representation. Nonetheless, these networks do not attempt to model a particular generative process, which requires an exponential number of hidden units and a large amount of training data to approximate an arbitrary binary distribution [18].

To capture nonlinearities or subpopulations, the distribution can be approximated using a finite *mixture* of Gaussians in which model estimation is often made tractable by working with a low-dimensional projection of the data (*e.g.,* [5, 9]). Nonetheless, this global projection often collapses or mixes the subpopulations, which derails learning the mixture structure of the underlying shape space [11]. Nonlinearity of shape statistics can also be modeled by lifting training shapes to a higher dimensional feature (aka kernel) space, where the shape distribution is often assumed to be Gaussian (*e.g.,* [15]). However, this approach results in an infinite-dimensional optimization scheme and sacrifices

the efficiency of optimizing in low-dimensional subspaces [9]. Further, one is then faced with finding the reverse mapping from feature space to shape space (aka preimage problem), which is often solved approximately.

In addressing the problem of modeling populations of label maps, this paper makes several contributions. *First*, it learns the generative model of a shape population directly in the label map space while not relying on ad hoc, implicit shape representations. *Second*, it proposes a *Bayesian* treatment of a latent variable model – ShapeOdds – as a low-dimensional *shape generating process*. *Third*, it extends recent works in the machine learning literature on variational bounds of logistic-Gaussian integrals designed to circumvent the intractable marginal likelihood and latent posterior leading to *deterministic* learning. *Fourth*, it presents a variational formulation that further reduces the sensitivity to hyperparameters by modeling posterior uncertainties. *Fifth*, it automatically estimates model hyperparameters – with closed-form re-estimation expressions – without the need for discrete searches and cross-validation. *Sixth*, it learns complex shape distributions directly in the high-dimensional shape space as a *mixture* of ShapeOdds models to balance the tradeoff between the model expressiveness with the efficiency of learning and inference without resorting to a global projection of training samples onto a low-dimensional linear subspace. By considering a probabilistic mixture of a latent variable model, we obtain a soft partitioning of the high-dimensional data space that requires only a few parameters for each mixture component and reduces the associated risk of overfitting. Experiments demonstrate that ShapeOdds and its mixture variant compare favorably with several different technologies for state-of-art shape modeling.

2 Latent Gaussian Model for Label Maps

Consider a raster defined over a spatial domain $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ containing D pixels. An object $\omega \subset \Omega$ is represented as a *label map* $\mathbf{f} \in \{0, 1\}^D$, where $\mathbf{f}(\mathbf{x}) = 1$, iff $\mathbf{x} \in \omega \;\; \forall \; \mathbf{x} \in \Omega$. In a *generative* sense, $\mathbf{f}$ is a realization of a *spatially correlated* field of D *Bernoulli* random variables defined on Ω with a pixelwise *parameter* $\mathbf{q}(\mathbf{x}) \in [0, 1]$ where $\mathbf{q}(\mathbf{x}) = p(\mathbf{x} \in \omega)$. Spatial regularity on the label maps, typically modeled as Markov random fields (MRFs), helps describe local correlations between nearby label values. The Bernoulli likelihood has an equivalent form in terms of the *exponential family distributions* that is parameterized by a field of real values $\boldsymbol{\phi}(\mathbf{x}) \in \mathbb{R}$, known as *natural parameters* (or LogOdds), where $\boldsymbol{\phi}(\mathbf{x}) = \text{logit}[\mathbf{q}(\mathbf{x})]$ with $\mathbf{q}(\mathbf{x})$ being the first moment of this form and hence denoted as *expectation* parameters. The merit in considering such an equivalence is casting any parameter estimation problem as an *unconstrained* optimization in the natural parameter space. Hence, the *generative model* of a label map includes a pixelwise Bernoulli likelihood and the MRF spatial prior.

$$p(\mathbf{f}|\boldsymbol{\phi}) = \left\{ \prod_{\mathbf{x} \in \Omega} \underbrace{\exp\left[\mathbf{f}(\mathbf{x})\boldsymbol{\phi}(\mathbf{x}) - \text{llp}[\boldsymbol{\phi}(\mathbf{x})]\right]}_{p(\mathbf{f}(\mathbf{x})|\boldsymbol{\phi}(\mathbf{x}))} \right\} \times \frac{1}{Z} \exp\left(-\frac{1}{T} U(\mathbf{f})\right) \tag{1}$$

where $\text{llp}[\phi] = \log\left(1 + e^{\phi}\right)$ is the *logistic-log-partition* function, $U(\mathbf{f})$ are clique potentials that favor spatially coherent silhouettes, Z is a Gibbs distribution

normalization constant, and T is its temperature. Learning a probability distribution over the label map space amounts to estimating the corresponding expectation parameters map $\mathbf{q}$ or equivalently the natural parameters (*i.e.,* LogOdds) map $\boldsymbol{\phi}$. Here, we coin the term *ShapeOdds* to refer to a data-driven estimation of *LogOdds* that places label maps in a vector space where global *shape* variation is taken into account in a statistically principled manner.

Consider an unknown *shape distribution* $p(\mathbf{f})$ in the label map space $\mathcal{F}$, of which we have only a finite ensemble $\mathrm{F} = \{\mathbf{f}_n\}_{n=1}^{N} \subset \mathcal{F}$. In the general case, let $p(\mathbf{f})$ be a multimodal distribution that is comprised of $K-$mixture components. In latent variable formalism, the distribution of the $k-$th component is governed by a low-dimensional *shape-generating process* – ShapeOdds – of L_k independent latent variables $\mathbf{z}^k \in \mathbb{R}^{L_k}$ where $L_k \ll D$. A label map $\mathbf{f}$ is also associated with a latent binary indicator $\mathbf{y} \in \{0,1\}^K$ that indicates the identity of the mixture component responsible for generating $\mathbf{f}$ where $\mathbf{1}_K^T\mathbf{y} = 1$. Here we consider a class of latent Gaussian models (LGMs) to capture correlations between observed voxels through Gaussian latent variables. In particular, the $k-$th component is drawn according to the mixing coefficients $p(\mathbf{y}^k = 1) = \pi_k$ where $\boldsymbol{\pi} \in [0,1]^K$ and $\mathbf{1}_K^T\boldsymbol{\pi} = 1$. A point $\mathbf{z}^k$ in the $k-$th *latent factor space* $\mathcal{Z}_k$ is then *generated* according to a Gaussian prior $p(\mathbf{z}^k) = \mathcal{N}(\mathbf{z}^k; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where $\boldsymbol{\mu}_k \in \mathbb{R}^{L_k}$ and $\boldsymbol{\Sigma}_k \in \mathbb{R}^{L_k \times L_k}$, which is mapped to $\boldsymbol{\phi}^k$ in the natural parameter space $\mathcal{P} \subset \mathbb{R}^D$ by a smooth mapping $h_k : \mathcal{Z}_k \to \mathcal{P}$. The logit function further maps $\mathcal{P}$ to the expectation parameter space $\mathcal{Q} \subset [0,1]^D$. A natural parameter map $\boldsymbol{\phi}^k \in \mathbb{R}^D$ is assumed to be confined to a linear subspace in $\mathcal{P}$ parameterized by a component-specific *factor loading* matrix $\mathbf{W}^k \in \mathbb{R}^{D \times L_k}$ and an *offset* vector $\mathbf{w}_0^k \in \mathbb{R}^D$ where $\boldsymbol{\phi}^k = h_k(\mathbf{z}^k) = \mathbf{W}^k\mathbf{z}^k + \mathbf{w}_0^k$. A $\boldsymbol{\phi}^k-$map thus induces a distribution $p(\mathbf{f}|\boldsymbol{\phi}^k)$ of label maps in $\mathcal{F}$. The corresponding $\boldsymbol{\phi}^k-$maps of F, although they lie in a linear subspace in $\mathcal{P}$, typically correspond to a nonlinear manifold in $\mathcal{F}$.

To alleviate overfitting and penalize highly irregular mappings h_k, we introduce a Gaussian MRF (GMRF) smoothness prior over individual loading/offset vectors $\{\mathbf{w}_l^k\}_{l=0}^{L}$ with $\mathbf{w}_l^k : \Omega \to \mathbb{R}^D$ where the prior on the mapping h_k can be factored out as $p(\mathbf{W}^k, \mathbf{w}_0^k) = \prod_{l=0}^{L_k} p(\mathbf{w}_l^k|\lambda_l^k)$. The smoothness prior over a vector $\mathbf{w}_l^k$ is expressed as a Gibbs distribution, $p(\mathbf{w}_l^k|\lambda_l^k) \propto \exp\left\{-\lambda_l^k E(\mathbf{w}_l^k)\right\}$ where $\lambda_l^k > 0$ is a hyperparameter that controls the generalizability aspect of the resultant mapping. The Gibbs energy $E(\mathbf{w}_l^k)$, hence, is chosen to favor smooth vectors by penalizing abrupt edges. We use Laplacian-square energy, *i.e.,* $E(\mathbf{w}_l^k) = \left\|\Delta\mathbf{w}_l^k\right\|_2^2$ to quantify the edges within $\mathbf{w}_l^k$.

The intrinsic dimensionality of the component-specific latent space $\mathcal{Z}_k$ is determined by the choice of L_k $\forall k \in \{1, ..., K\}$. Nonetheless, an exhaustive grid search over this choice becomes computationally intractable in mixture modeling, because each component would have a different L_k. The probabilistic formulation of LGMs allows this discrete model selection to be handled within the Bayesian paradigm. We make use of the sparsity-inducing automatic relevance determination (ARD) prior to further regularize the solution space via a parameterized data-driven prior distribution that effectively prunes away irrelevant factors of variations [19]. We introduce an ARD prior on the loading vectors $\{\mathbf{w}_l^k\}_{l=1}^{L_k}$ with

L_k set to the maximum allowed dimensionality, *i.e.,* $L_k = N - 1$ and $N \ll D$. ARD is a zero-mean isotropic Gaussian prior parameterized by $\beta_l^k \in \mathbb{R}_{>0}$ such that $p(\mathbf{w}_l^k|\beta_l^k) = \mathcal{N}(\mathbf{w}_l^k; \mathbf{0}_D, (\beta_l^k)^{-1}\mathbf{I}_D)$ where $\mathbf{0}_D$ and $\mathbf{I}_D$ are the zero vector and identity matrix in $\mathbb{R}^D$, respectively. During the learning process, $\beta_l^k \to \infty$ for irrelevant factors to remove the unnecessary complexity of the resulting model.

Thus, a mixture of *ShapeOdds* refers to the shape-generating process with model parameters $\Theta = \{\Theta_k\}_{k=1}^K$ and priors hyperparameters $\Psi = \{\Psi_k\}_{k=1}^K$. The parameters of the $k-$th ShapeOdds are $\Theta_k = \{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \mathbf{W}^k, \mathbf{w}_0^k\}$, and its hyperparameters are $\Psi_k = \{\boldsymbol{\lambda}^k, \boldsymbol{\beta}^k\}$ where $\boldsymbol{\lambda}^k \in \mathbb{R}_{>0}^{L_k+1}$ and $\boldsymbol{\beta}^k \in \mathbb{R}_{>0}^{L_k}$. The underlying generative process can be defined as follows:

$$\mathbf{y}_n|\Theta \sim \text{Cat}[K, \boldsymbol{\pi}], \quad \mathbf{z}_n^k|\Theta_k \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad \mathbf{f}_n|\mathbf{z}_n^k, \Theta_k \sim \text{Expon}[\boldsymbol{\phi}_n^k]\,\text{Mrf}[\nu] \quad (2)$$

$$\text{Expon}[\boldsymbol{\phi}_n^k] \doteq \prod\nolimits_{\mathbf{x}\in\Omega} \text{Expon}[\boldsymbol{\phi}_n^k(\mathbf{x})], \quad \boldsymbol{\phi}_n^k = \mathbf{W}^k\mathbf{z}_n^k + \mathbf{w}_0^k, \quad (3)$$

$$\mathbf{w}_0^k|\lambda_0^k \sim \text{GMrf}[\lambda_0^k], \quad \mathbf{w}_l^k|\lambda_l^k, \beta_l^k \sim \text{GMrf}[\lambda_l^k]\,\text{Ard}[\beta_l^k] \quad (4)$$

with $\text{GMrf}[\lambda] \doteq \mathcal{N}(\mathbf{0}_D, \lambda^{-1}\mathbf{S})$ where $\mathbf{S}$ is the structure matrix containing the stencil of the negative bi-Laplacian operator; the first variation of $E(\mathbf{w})$. The MRF prior in (2), whose hyperparameter $\nu > 0$ is related to the temperature in (1), reflects the spatial regularity of the given label maps. Note that the choice of ν does not affect the model learning process. Eq (3) is due to the axiom of local/conditional independence, *i.e.,* the observed variables are conditionally independent given the latent variables, where $\text{Expon}[\boldsymbol{\phi}(\mathbf{x})]$ is given by (1).

3 Model Variational Learning

Consider a set of *i.i.d.* label maps $\text{F} = \{\mathbf{f}_n\}_{n=1}^N$. The corresponding *discrete* latent indicator variables are denoted as $\text{Y} = \{\mathbf{y}_n\}_{n=1}^N$. The component-specific *continuous* latent factor variables $\text{Z}_k = \{\mathbf{z}_n^k\}_{n=1}^N$ define the low-dimensional representation w.r.t. the $k-$th generating mixture component. The *marginal likelihood* of F under the mixture model can be formulated as a marginalization over the latent indicator and factor variables.

$$p(\text{F}|\Theta) = \prod\nolimits_{n=1}^N \sum\nolimits_{k=1}^K \int_{\mathcal{Z}_k} p(\mathbf{f}_n|\mathbf{z}_n^k, \Theta_k) p(\mathbf{z}_n^k|\Theta_k) p(\mathbf{y}_n^k = 1|\Theta) d\mathbf{z}_n^k \quad (5)$$

Let $q(\mathbf{y}_n^k = 1, \mathbf{z}_n^k) = q(\mathbf{z}_n^k|\mathbf{y}_n^k = 1)q(\mathbf{y}_n^k = 1)$ be the approximate variational distribution of $k-$th component latent variables. The *lower bound* to the log-marginal likelihood can thus be obtained by dividing and multiplying by the posterior approximate and then applying the Jensen inequality.

$$\mathcal{L}(\Theta) \geq \underline{\mathcal{L}}^J(\Theta) = \sum\nolimits_{k=1}^K q(\mathbf{y}_n^k = 1) \left\{ \log \frac{p(\mathbf{y}_n^k = 1|\Theta)}{q(\mathbf{y}_n^k = 1)} \right.$$

$$\left. + \sum\nolimits_{n=1}^N \left(\mathbb{E}_{q(\mathbf{z}_n^k|\mathbf{y}_n^k=1)} \left[\log \frac{p(\mathbf{z}_n^k|\Theta_k)}{q(\mathbf{z}_n^k|\mathbf{y}_n^k = 1)} \right] + \mathbb{E}_{q(\mathbf{z}_n^k|\mathbf{y}_n^k=1)} \left[\log p(\mathbf{f}_n|\mathbf{z}_n, \Theta) \right] \right) \right\} \quad (6)$$

The variational distributions $q(.)$ that would tighten the Jensen lower bound in (6) should satisfy $\partial\underline{\mathcal{L}}^J(\Theta)/\partial q(.) = 0$. Under variational Bayes with exponential family distribution data likelihood, the variational latent factor posterior follows the same family as the factor prior, *i.e.,* Gaussian, leading to a tractable integral. Hence, $q(\mathbf{z}_n^k|\boldsymbol{\gamma}_n^k) = \mathcal{N}(\mathbf{z}_n^k|\mathbf{m}_n^k, \mathbf{V}_n^k)$, where $\boldsymbol{\gamma}_n^k = \{\mathbf{m}_n^k, \mathbf{V}_n^k\}$, is a Gaussian approximate to the factor posterior with mean $\mathbf{m}_n^k \in \mathbb{R}^{L_k}$ and covariance $\mathbf{V}_n^k \in \mathbb{R}^{L_k \times L_k}$.

The *first expectation* term in (6) is the negative KullbackLeibler (KL) divergence that pushes the variational posterior $q(\mathbf{z}_n^k|\mathbf{y}_n^k = 1)$ to the respective Gaussian prior $p(\mathbf{z}_n^k|\Theta_k)$. Its closed form is given by

$$-\mathrm{KL}_{\mathbf{z}_n^k|\mathbf{y}_n^k} = \frac{1}{2}\left\{\log|\mathbf{V}_n^k\boldsymbol{\Sigma}_k^{-1}| - \mathrm{Tr}[\mathbf{V}_n^k\boldsymbol{\Sigma}_k^{-1}] - (\mathbf{m}_n^k - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{m}_n^k - \boldsymbol{\mu}_k) + L_k\right\}$$

Taking the functional derivative of $\underline{\mathcal{L}}^J(\Theta)$ w.r.t. $q(\mathbf{y}_n^k = 1)$ and equating to zero, the variational posterior for the latent indicator variables can be written as

$$\begin{aligned}\log q(\mathbf{y}_n^k = 1) = \log p(\mathbf{y}_n^k = 1|\Theta) + \mathbb{E}_{q(\mathbf{z}_n^k|\mathbf{y}_n^k=1)}\left[\log\left\{p(\mathbf{z}_n^k|\Theta_k)/q(\mathbf{z}_n^k|\mathbf{y}_n^k = 1)\right\}\right] \\ + \mathbb{E}_{q(\mathbf{z}_n^k|\mathbf{y}_n^k=1)}\left[\log p(\mathbf{f}_n|\mathbf{z}_n, \Theta)\right] + \text{const} \qquad (7)\end{aligned}$$

The distribution in (7) does not admit to a well-known distribution of a closed-form especially with the second expectation term, which is not tractable. Nonetheless, we only care about the relative aspect of the indicator variables. Hence, we typically compute $\log q(\mathbf{y}_n^k = 1)$ based on the current model and factor variational parameters and then normalize the distribution such that $\sum_{k=1}^K q(\mathbf{y}_n^k = 1) = 1$. The first term in (6) is thus the negative KL divergence that pushes the indicator variational posterior to the respective component prior, *i.e.,* $-\mathrm{KL}_{\mathbf{y}_n^k} = \log\pi_k - \log q(\mathbf{y}_n^k = 1)$ where $p(\mathbf{y}_n^k = 1|\Theta) = \pi_k$.

The *second expectation* term in (6) can be expressed in the natural parameter space $\mathcal{P}$ according to the mapping $h_k(\mathbf{z}^k)$ and the conditional independence in (3), in which the Gaussian approximate posterior $q(\mathbf{z}_n^k|\mathbf{y}_n^k = 1)$ in $\mathcal{Z}_k$ induces a per-pixel Gaussian posterior $q(\boldsymbol{\phi}_n^k(\mathbf{x})|\widetilde{\boldsymbol{\gamma}}_n^k(\mathbf{x}))$ in $\mathcal{P}$ with $\widetilde{\boldsymbol{\gamma}}_n^k(\mathbf{x}) = \{\widetilde{\mathbf{m}}_n^k(\mathbf{x}), \widetilde{\mathbf{V}}_n^k(\mathbf{x})\}$ where $\mathbf{w}_0^k(\mathbf{x}) \in \mathbb{R}$ and $\mathbf{W}^k(\mathbf{x}) \in \mathbb{R}^{L_k}$.

$$\widetilde{\mathbf{m}}_n^k(\mathbf{x}) = \mathbf{W}^k(\mathbf{x})\mathbf{m}_n^k + \mathbf{w}_0^k(\mathbf{x}), \quad \widetilde{\mathbf{V}}_n^k(\mathbf{x}) = \mathbf{W}^k(\mathbf{x})\mathbf{V}_n^k[\mathbf{W}^k(\mathbf{x})]^T \qquad (8)$$

Note that the spatial coherency is still promoted through the GMRF prior on the offset and loading vectors. Using the exponential form of the Bernoulli likelihood in (1), the *second expectation* term in (6) can be expressed in $\mathcal{P}$ as

$$\sum\nolimits_{\mathbf{x}\in\Omega}\left\{\mathbf{f}_n(\mathbf{x})\widetilde{\mathbf{m}}_n^k(\mathbf{x}) - \mathbb{E}_{q(\boldsymbol{\phi}_n^k(\mathbf{x})|\widetilde{\boldsymbol{\gamma}}_n^k(\mathbf{x}))}\left[\mathrm{llp}[\boldsymbol{\phi}_n^k(\mathbf{x})]\right]\right\} \qquad (9)$$

$$\geq \sum\nolimits_{\mathbf{x}\in\Omega}\underline{\mathcal{B}}_n^k(\mathbf{x}) := \left\{\mathbf{f}_n(\mathbf{x})\widetilde{\mathbf{m}}_n^k(\mathbf{x}) - \overline{\mathcal{B}}(\widetilde{\boldsymbol{\gamma}}_n^k(\mathbf{x}), \boldsymbol{\alpha}_n^k(\mathbf{x}))\right\} \qquad (10)$$

Eq (9) is intractable due to the llp function and can be lower-bounded in $\mathcal{P}-$space by defining an upper bound $\overline{\mathcal{B}}$ for the expectation of the llp function with *local*, *i.e.,* per-pixel, variational parameters $\boldsymbol{\alpha}_n^k(\mathbf{x})$. To avoid the recomputation of per-pixel/per-component/per-sample $\boldsymbol{\alpha}_n^k(\mathbf{x})$, we use a *fixed* piecewise quadratic upper bound for the llp function recently proposed in [20] as a proven tight bound compared to other quadratic bounds, where $\boldsymbol{\alpha}_n^k(\mathbf{x}) = \boldsymbol{\alpha}\ \forall\ n, k, \mathbf{x}$ are optimized in advance via a minimax optimization to ensure a tight bound. The upper bound $\overline{\mathcal{B}}$ can thus be expressed in terms of truncated Gaussian moments, due to the approximate Gaussian posterior, whose closed-form expressions along with their gradients are available [20]. The new bound reads as

$$\underline{\mathcal{L}}(\Theta, \boldsymbol{\gamma}, \boldsymbol{\alpha}) = \sum\nolimits_{n=1}^N\sum\nolimits_{k=1}^K \underbrace{q(\mathbf{y}_n^k = 1)\left\{-\mathrm{KL}_{\mathbf{y}_n^k} - \mathrm{KL}_{\mathbf{z}_n^k|\mathbf{y}_n^k} + \sum\nolimits_{\mathbf{x}\in\Omega}\underline{\mathcal{B}}_n^k(\mathbf{x})\right\}}_{\underline{\mathcal{L}}_n^k(\Theta_k, \boldsymbol{\gamma}_n^k, \boldsymbol{\alpha})} \qquad (11)$$

We propose a variational expectation-maximization (EM) algorithm that optimizes the rigorous lower bound defined in (11). The E-step in (13) optimizes

the variational indicator and factor posterior distributions at an iteration i given the current guess of model parameters $\Theta^{(i-1)}$. The M-step chooses the next guess of $\Theta^{(i)}$ to maximize the *regularized* variational bound in (17). Iterating between these two steps involves concave optimizations due to the concavity of the lower bound in (11) [20] and the semipositive definiteness of the bi-Laplacian operator, for which we can use gradient-based optimization (see Algorithm 1 for gradient expressions). The maximum-a-posteriori (MAP) objective of the offset and loading vectors, after removing constant terms, can be written as in (22). The first variation of (22) w.r.t. the offset and loading vectors reads as in (23) and (24): the vectors $\mathbf{g}_n^{\mathbf{m},k}$ and $\mathbf{G}_n^{\mathbf{V},k}$ are bound gradients in (16), $\odot$ refers to a Hadamard product, $\mathbf{m}_{n,l}^k$ is the $l-$th entry of $\mathbf{m}_n^k$, and $\mathbf{V}_{n,l}^k$ is the $l-$th column of $\mathbf{V}_n^k$. To enable large time steps Δt while maintaining stable updates, we use a *semi-implicit scheme* with finite-forward time marching to define iterative updates for $\mathbf{w}_l^k$'s in (26), where spatial convolution $\otimes$ can be efficiently performed as multiplication in the Fourier domain.

Hyperparameters: To complete our Bayesian treatment, we formulate an *evidence approximation*, aka type-II maximum likelihood, in which we marginalize over Ψ_k for each mixture component. We consider a Jeffrey prior to construct a hyper-hyperparameters-free noninformative hyperprior on the hyperparameters, leading to analytic integrals. The noninformative hyperpriors on λ_l^k and β_l^k are $p(\lambda_l^k) \propto 1/\lambda_l^k$ and $p(\beta_l^k) \propto 1/\beta_l^k$, respectively. The marginalization over the hyperparameters involves λ_l^k-integrals and β_l^k-integrals, each with an analytic form $\frac{\Gamma(D/2)|\mathbf{S}|^{1/2}}{\pi^{D/2}[(\mathbf{w}_l^k)^T\mathbf{S}\mathbf{w}_l^k]^{D/2}}$ and $\frac{\Gamma(D/2)|\mathbf{S}|^{1/2}}{\pi^{D/2}[(\mathbf{w}_l^k)^T\mathbf{w}_l^k]^{D/2}}$, respectively, using a $\Gamma-$function integral form. For a given Ψ_k, the gradient of $\log p(\Theta_k|\Psi_k)$ w.r.t. $\mathbf{w}_l^k$ should coincide with that of the marginal $p(\Theta_k)$ [21]. Hence, the effective values of the hyperparameters in (30) re-estimate Ψ_k after each pair of E- and M-steps to compute the new Ψ_k given the current guess of Θ_k.

4 Results

We assessed the *generalization* of ShapeOdds and its mixture variant – compared to *baseline* models – w.r.t. modeling unseen shapes, which reflect the ability of a learned density function to spread out between and around the training shapes.

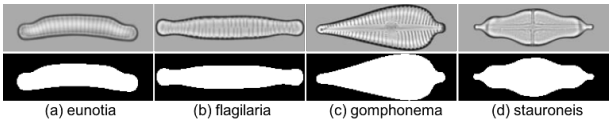

Fig. 1. Diatoms dataset: (top) diatom images of the four classes and (bottom) the corresponding label maps.

Datasets: We considered two datasets. (1) The *diatoms* dataset[1] contains 360 images from four diatom classes – eunotia, flagilaria, gomphonema, and stauroneis (see Fig. 1) – collected as part of the automatic diatom identification and classification project [22]. This dataset manifests learning ShapeOdds in a high-dimensional shape space that possesses subpopulations with limited training samples compared to significant shape variations across successive generations in a clone, thereby constructing nonlinear shape manifolds. (2) The cardiac *left*

[1] Downloaded from the DIADIST project page: rbg-web2.rbge.org.uk/DIADIST/

Algorithm 1 Variational EM for learning mixture of ShapeOdds

E-Step: Optimize variational parameters given the current model parameters.

(a) Compute the variational distributions of the indicator variables, *i.e.,* $q(\mathbf{y}_n^k = 1),\ \forall n \in \{1, .., N\}$ and $k \in \{1, ..., K\}$ such that $\sum_{k=1}^{K} q(\mathbf{y}_n^k = 1) = 1$ using $\Theta^{(i-1)}$ and $\boldsymbol{\gamma}_n^{k\,(i-1)}$ where

$$\log q(\mathbf{y}_n^k = 1) = \pi_k^{(i-1)} - \mathrm{KL}_{\mathbf{z}_n^k|\mathbf{y}_n^k} + \sum\nolimits_{\mathbf{x}\in\Omega} \underline{\mathcal{B}}_n^k(\mathbf{x}) \tag{12}$$

(b) Optimize the variational parameters of the latent factors approximate posterior, *i.e.,* $q(\mathbf{z}_n^k|\mathbf{y}_n^k = 1),\ \forall n \in \{1, .., N\}$ and $k \in \{1, ..., K\}$.

$$\boldsymbol{\gamma}_n^{k\,(i)} = \mathrm{argmax}_{\gamma_n^k}\, \underline{\mathcal{L}}_n^k(\Theta_k^{(i-1)}, \boldsymbol{\gamma}_n^k, \boldsymbol{\alpha}) \tag{13}$$

$$\frac{\partial \underline{\mathcal{L}}_n^k}{\partial \mathbf{m}_n^k} = q(\mathbf{y}_n^k = 1)\left\{-\boldsymbol{\Sigma}_k^{-1}(\mathbf{m}_n^k - \boldsymbol{\mu}_k) + \sum\nolimits_{\mathbf{x}\in\Omega} g_n^{\mathbf{m},k}(\mathbf{x})[\mathbf{W}^k(\mathbf{x})]^T\right\} \tag{14}$$

$$\frac{\partial \underline{\mathcal{L}}_n^k}{\partial \mathbf{V}_n^k} = q(\mathbf{y}_n^k = 1)\left\{\frac{1}{2}\left[(\mathbf{V}_n^k)^{-1} - \boldsymbol{\Sigma}_k^{-1}\right] + \sum\nolimits_{\mathbf{x}\in\Omega} G_n^{\mathbf{V},k}(\mathbf{x})[\mathbf{W}^k(\mathbf{x})]^T\mathbf{W}^k(\mathbf{x})\right\} \tag{15}$$

$$\text{where}\quad g_n^{\mathbf{m},k}(\mathbf{x}) = \frac{\partial \underline{\mathcal{B}}_n^k(\mathbf{x})}{\partial \widetilde{\mathbf{m}}_n^k(\mathbf{x})},\quad G_n^{\mathbf{V},k}(\mathbf{x}) = \frac{\partial \underline{\mathcal{B}}_n^k(\mathbf{x})}{\partial \widetilde{\mathbf{V}}_n^k(\mathbf{x})} \tag{16}$$

M-Step: Optimize model parameters given the current variational parameters.

$$\Theta^{(i)} = \arg\max_{\Theta}\left\{\sum\nolimits_{n=1}^{N}\sum\nolimits_{k=1}^{K} \underline{\mathcal{L}}_n^k(\Theta_k, \boldsymbol{\gamma}_n^{k\,(i)}, \boldsymbol{\alpha})\right\} + \log p(\Theta|\Psi) \tag{17}$$

$$\text{where}\quad p(\Theta|\Psi) = \prod\nolimits_{k=1}^{K} p(\mathbf{w}_0^k|\lambda_0^k)\prod\nolimits_{l=1}^{L_k} p(\mathbf{w}_l^k|\lambda_l^k)p(\mathbf{w}_l^k|\beta_l^k) \tag{18}$$

(a) Compute component-specific latent factor prior parameters.

$$\pi_k = \frac{1}{N}\sum\nolimits_{n=1}^{N} q(\mathbf{y}_n^k = 1) \tag{19}$$

$$\boldsymbol{\mu}_k = \frac{1}{N_k}\sum\nolimits_{n=1}^{N} q(\mathbf{y}_n^k = 1)\mathbf{m}_n^k \quad \text{where} \quad N_k = \sum\nolimits_{n=1}^{N} q(\mathbf{y}_n^k = 1) \tag{20}$$

$$\boldsymbol{\Sigma}_k = \frac{1}{N_k}\sum\nolimits_{n=1}^{N} q(\mathbf{y}_n^k = 1)\left\{\mathbf{V}_n^k + (\mathbf{m}_n^k - \boldsymbol{\mu}_k)(\mathbf{m}_n^k - \boldsymbol{\mu}_k)^T\right\} \tag{21}$$

(b) Optimize component-specific factor model $\mathbf{W}^k, \mathbf{w}_0^k\ \forall k = \{1, ..., K\}$.

$$\mathcal{E}(\Theta|\boldsymbol{\gamma}, \boldsymbol{\alpha}, \Psi) = -\underline{\mathcal{L}}(\Theta, \gamma, \boldsymbol{\alpha}) + \sum\nolimits_{k=1}^{K}\left\{\left(\lambda_0^k/2\right)(\mathbf{w}_0^k)^T\mathbf{S}\mathbf{w}_0^k + \sum\nolimits_{l=1}^{L_k}\left[\left(\lambda_l^k/2\right)(\mathbf{w}_l^k)^T\mathbf{S}\mathbf{w}_l^k + \left(\beta_l^k/2\right)(\mathbf{w}_l^k)^T\mathbf{w}_l^k\right]\right\} \tag{22}$$

$$\frac{d\mathcal{E}}{d\mathbf{w}_0^k} = -\sum\nolimits_{n=1}^{N} q(\mathbf{y}_n^k = 1)\ \mathbf{g}_n^{\mathbf{m},k} + \lambda_0^k\mathbf{S}\mathbf{w}_0^k, \tag{23}$$

$$\frac{d\mathcal{E}}{d\mathbf{w}_l^k} = -\left\{\frac{\partial \underline{\mathcal{L}}}{\partial \mathbf{w}_l^k} + \lambda_l^k\mathbf{S}\mathbf{w}_l^k + \beta_l^k\mathbf{w}_l^k\right\} \tag{24}$$

$$\frac{\partial \underline{\mathcal{L}}}{\partial \mathbf{w}_l^k} = \sum\nolimits_{n=1}^{N} q(\mathbf{y}_n^k = 1)\Big[\mathbf{g}_n^{\mathbf{m},k}\mathbf{m}_{n,l}^k + 2\mathbf{G}_n^{\mathbf{V},k} \odot \mathbf{W}^k\mathbf{V}_{n,l}^k\Big] \tag{25}$$

$$\mathbf{w}_l^{k\,(t)} = \left\{\frac{1}{1 + \Delta t\lambda_l^k\mathbf{S}}\right\} \otimes \left\{\mathbf{w}_l^{k\,(t-1)} + \Delta t\left[\frac{\partial \underline{\mathcal{L}}}{\partial \mathbf{w}_l^k} - \delta(l > 0)\beta_l^k\mathbf{w}_l^{k\,(t-1)}\right]\right\} \tag{26}$$

H-Step:

$$p(\Theta) = \prod_{k=1}^{K}\prod_{l=0}^{L_k}\left\{\int_0^\infty p(\mathbf{w}_l^k|\lambda_l^k)p(\lambda_l^k)d\lambda_l^k\right\} \times \prod_{k=1}^{K}\prod_{l=1}^{L_k}\left\{\int_0^\infty p(\mathbf{w}_l^k|\beta_l^k)p(\beta_l^k)d\beta_l^k\right\} \tag{27}$$

$$p(\mathbf{w}_l^k|\lambda_l^k) = \frac{(\lambda_l^k)^{D/2}|\mathbf{S}|^{1/2}}{(2\pi)^{D/2}}\exp\left\{-\frac{\lambda_l^k}{2}(\mathbf{w}_l^k)^T\mathbf{S}\mathbf{w}_l^k\right\} \tag{28}$$

$$p(\mathbf{w}_l^k|\beta_l^k) = \frac{(\beta_l^k)^{D/2}}{(2\pi)^{D/2}}\exp\left\{-\frac{\beta_l^k}{2}(\mathbf{w}_l^k)^T\mathbf{w}_l^k\right\} \tag{29}$$

$$\lambda_l^k = \frac{D}{(\mathbf{w}_l^k)^T\mathbf{S}\mathbf{w}_l^k},\ \beta_l^k = \frac{D}{(\mathbf{w}_l^k)^T\mathbf{w}_l^k}\ \text{ where }\ (\mathbf{w}_l^k)^T\mathbf{S}\mathbf{w}_l^k = ||\Delta\mathbf{w}_l^k||^2 \tag{30}$$

atrial (LA) shape dataset[2] contains 60 late-gadolinium enhancement (LGE) MRI volumes with an isotropic resolution of 1.25mm and expert-delineated binary segmentations for the epicardium. This dataset is an example of 3D anatomical shapes with relatively small training sample size that exhibit high variability, especially within and around the pulmonary veins (see Fig. 2).

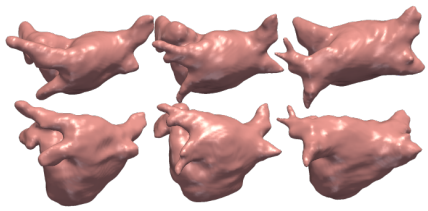

Fig. 2. LA training samples: (top) superior and (bottom) posterior-anterior views.

Baseline models: For comparison, we considered the state-of-the-art ShapeBM [17], which learns shape models directly in the label map space. For diatoms, we trained ShapeBM with 1×6 receptive fields and an overlap of 20 pixels using pre-training and 1000 epochs. We used 1800 and 200 hidden units for the first and second layers, respectively. We also learned shape models using PCA directly in the label map space (PCA-Prob-LM) as in [14]. We further considered current practices that use intermediate embeddings such as signed distance maps (SDMs) and Gaussian smoothed label maps (GAUSS), where the multiplicative factor for SDMs and the width of the Gaussian kernel were optimized using cross-validation. We learned shape models using PCA in the LogOdds space, similar to [7], and in the expectation parameter space (Prob), similar to [8]. For mixture modeling of diatoms and LA datasets, we used a mixture of probabilistic PCA [23] being learned in the LogOdds space using SDMs- and GAUSS-based representations in which we parameterized the component-specific covariance matrix through its eigen decomposition. For nonparametric models, we focused on the kernel density estimate (KDE) using SDMs as in [6] where we fixed the kernel width to be the mean squared nearest-neighbor distance.

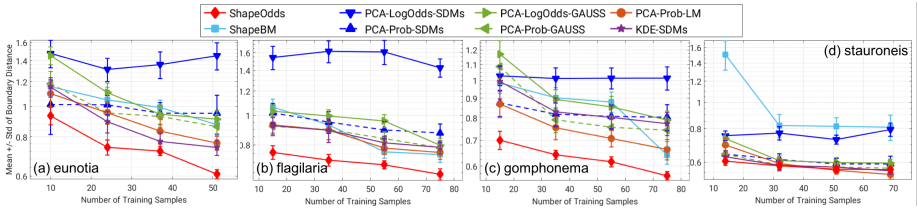

Fig. 3. Statistics of models' generalization on each diatom class. Lower is better.

Evaluation: We used a boundary-based evaluation metric to provide better insight – compared to overlap-based metrics [24] – into how close an estimated $\mathbf{q}-$map is to its label map. In particular, we used average boundary distance [24] among all pairwise boundary distances (in pixels and mm for diatoms and LA datasets, respectively) between the $0.5-$crossing of the groundtruth label map and the $0.5-$levelset of the estimated $\mathbf{q}-$map.

Proof-of-concept: We considered each diatom class individually. Fig. 3 demonstrates the generalization performance of ShapeOdds compared to baseline mod-

[2] Downloaded from the national alliance for medical image computing (NAMIC) project page: https://www.na-mic.org/Wiki/index.php/DBP3:Utah

els as a function of the training sample size where training subsets of $N = \{15\%, 35\%, 55\%, 75\%\}$ of the full dataset were randomly drawn five times.

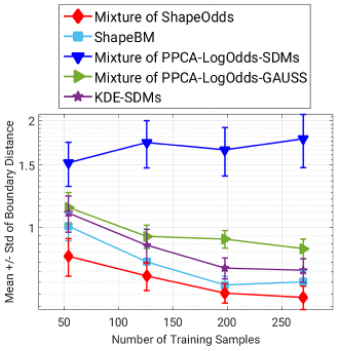

Fig. 4. Generalization statistics on all diatom classes using mixtures of four components.

The poor performance of models that have been learned in the LogOdds space is evident regardless the deployed intermediate representation compared to the expectation parameter space. This reveals that existing ad hoc embeddings are not statistically principled approaches to embed label maps in the natural parameter space. The poorer performance of SDMs-based models indicates that they lead to suboptimal generative models that do not generalize well on unseen data. In particular, signed distance to the shape's boundary is a geometric representation that does not correlate well with the underlying generative process. GAUSS-based models make use of more training samples for better generalization. However, simply blurring the label maps smoothes out shape features in a way that is blind to the underlying shape variability in the input data. KDE-SDMs construct a non-parametric estimate of the expectation parameters based on the similarity of the unseen shape's SDM to each training sample. KDE-SDMs shows a slightly better performance for eunotia and stauroneis classes; however, it tends to overfit with the other *highly variable* classes. This mode of failure is typical for this model and appears to be an inability to find a good set of weights on training samples to, in turn, recover a good parameter map. ShapeBM advocates an axis-aligned shape space partitioning, in a non-data-driven manner, with a weight-sharing scheme to balance the number of parameters to estimate and the generality of the model. This explains its drastic performance with the stauroneis class where no unique ideal network architecture can be used for all datasets. On the other hand, ShapeOdds compares favorably against all baseline models and shows better generalization performance even with small training sizes compared to the underlying variability.

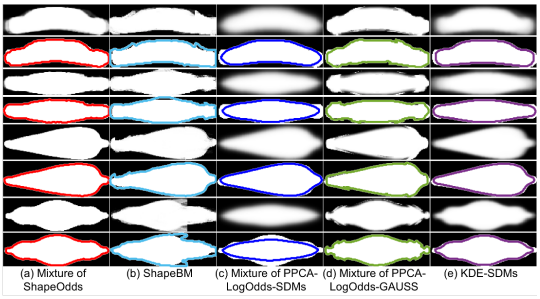

Fig. 5. Mixture results: (odd rows) samples of estimated $\mathbf{q}-$maps and (even rows) 0.5$-$levelset of the $\mathbf{q}-$map overlayed on the groundtruth label map.

Mixture experiment: Here we considered samples from all the diatom classes. Fig. 4 depicts the generalization performance as a function of the training sample size. One can note the poor generalization of the SDMs-based mixture model compared to the GAUSS-based one, which emphasizes the inadequacy of a geometric representation to learn a probability distribution over the label maps.

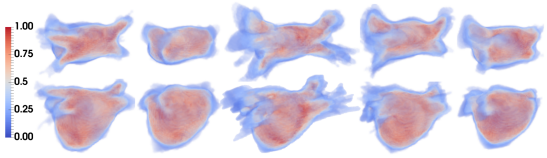

Fig. 6. Volume rendering (top) superior and (bottom) posterior-anterior views of the $\mathbf{q}-$maps of ShapeOdds mixture components means.

Extending ShapeOdds to mixtures achieves the balance between model capacity and generalizability in a *data-driven* manner and a statistically principled approach – rather than heuristically partitioning the shape space as in ShapeBM – by parameterizing each mixture component by its dominant subspace. Fig. 5 demonstrates that the mixture of ShapeOdds generalizes to unseen examples with crisp $\mathbf{q}-$maps compared to other baseline models. One can notice the tendency of SDMs-based models (KDE and mixture) to recover over-smoothed $q-$maps, revealing a failure to learn enough shape variability and leading to parameter maps that do not preserve shape class features such as different shapes of diatom valves. The effect of ShapeBM's space partitioning can be observed where the lack of *enough* receptive field overlap leads to discontinuities in the reconstructed $\mathbf{q}-$map and fails to perserve global shape features.

For the LA shapes, training subsets of $N = 75\%$ of the full dataset were randomly drawn three times to train mixture models of five components. Fig. 6 visualizes the corresponding expectation parameters of the ShapeOdds mixture means, *i.e.*, $\mathbf{q}_0^k = \text{logit}[\mathbf{w}_0^k]$. The significant LA shape variabilty is evident by having representative parameter maps that pertain to different shape characteristics, especially w.r.t. the elongation and curvature of the pulmonary veins. Table 1 reports the mean and std of the average boundary distance between unseen LA samples and estimated $\mathbf{q}-$maps from different models. Fig. 7 shows such a boundary distance on samples of epicardial surfaces where the majority of deviation is concentrated around the pulmonary veins, a highly variable LA shape feature. The mixture of ShapeOdds is able to estimate $\mathbf{q}-$maps with veins that preserve better proximity to the groundtruth shapes compared to other models.

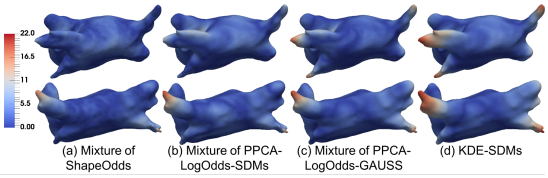

Fig. 7. Boundary distances as color maps on sample epicardial meshes (superior view).

Table 1. Statistics of models generalization on the LA dataset (Lower is better)

Mixture of ShapeOdds	Mixture of PPCA-LogOdds-SDMs	Mixture of PPCA-LogOdds-GAUSS	KDE-SDMs
$\mathbf{1.83 \pm 0.23}$ mm	1.94 ± 0.3 mm	2.05 ± 0.45 mm	2.47 ± 0.67 mm

5 Conclusion and Future Work

We presented a probabilistic generative shape model that offers a data-driven placement of label maps in a vector space without relying on any intermediate representation. Experiments have shown that the proposed model generalizes to unseen samples in nontrivial ways. In the future, we plan to pursue deep latent models to allow scaling to higher resolution shapes while avoiding overfitting.

References

1. Heimann, T., Meinzer, H.P.: Statistical shape models for 3d medical image segmentation: a review. MedIA 13(4), 543–563 (2009)
2. Bredbenner, T.L., Eliason, T.D., Potter, R.S., Mason, R.L., Havill, L.M., Nicolella, D.P.: Statistical shape modeling describes variation in tibia and femur surface geometry between control and incidence groups from the osteoarthritis initiative database. JBiomech 43(9), 1780–1786 (2010)
3. Balestra, S., Schumann, S., Heverhagen, J., Nolte, L., Zheng, G.: Articulated statistical shape model-based 2d-3d reconstruction of a hip joint. In: IPCAI. pp. 128–137. Springer (2014)
4. Shen, K.k., Fripp, J., Mériaudeau, F., Chételat, G., Salvado, O., Bourgeat, P.: Detecting global and local hippocampal shape changes in alzheimer's disease using statistical shape models. Neuroimage 59(3), 2155–2166 (2012)
5. Rousson, M., Paragios, N., Deriche, R.: Implicit active shape models for 3d segmentation in mr imaging. In: MICCAI, pp. 209–216. Springer (2004)
6. Cremers, D., Osher, S.J., Soatto, S.: Kernel density estimation and intrinsic alignment for shape priors in level set segmentation. IJCV 69(3), 335–351 (2006)
7. Pohl, K.M., Fisher, J., Bouix, S., Shenton, M., McCarley, R.W., Grimson, W.E.L., Kikinis, R., Wells, W.M.: Using the logarithm of odds to define a vector space on probabilistic atlases. MedIA 11(5), 465–477 (2007)
8. Cremers, D., Schmidt, F.R., Barthel, F.: Shape priors in variational image segmentation: Convexity, lipschitz continuity and globally optimal solutions. In: CVPR. pp. 1–6. IEEE (2008)
9. Rousson, M., Cremers, D.: Efficient kernel density estimation of shape and intensity priors for level set segmentation. In: MICCAI, pp. 757–764. Springer (2005)
10. Wimmer, A., Soza, G., Hornegger, J.: A generic probabilistic active shape model for organ segmentation. In: MICCAI, pp. 26–33. Springer (2009)
11. Dasgupta, S.: Learning mixtures of gaussians. In: FCS. pp. 634–644. IEEE (1999)
12. Ashburner, J., Friston, K.J.: Computing average shaped tissue probability templates. Neuroimage 45(2), 333–341 (2009)
13. Sabuncu, M.R., Yeo, B.T., Van Leemput, K., Fischl, B., Golland, P.: A generative model for image segmentation based on label fusion. TMI 29(10), 1714–1729 (2010)
14. Dambreville, S., Rathi, Y., Tannenbaum, A.: A shape-based approach to robust image segmentation. In: ICIAR. pp. 173–183. Springer (2006)
15. Dambreville, S., Rathi, Y., Tannenbaum, A.: A framework for image segmentation using shape models and kernel space shape priors. TPAMI 30(8), 1385–1399 (2008)
16. Welling, M., Chemudugunta, C., Sutter, N.: Deterministic latent variable models and their pitfalls. SIAM-DM pp. 196–207 (2008)
17. Eslami, S.A., Heess, N., Williams, C.K., Winn, J.: The shape boltzmann machine: a strong model of object shape. IJCV 107(2), 155–176 (2014)
18. Freund, Y., Haussler, D.: Unsupervised learning of distributions on binary vectors using two layer networks. In: NIPS. pp. 912–919 (1992)
19. Wipf, D.P., Nagarajan, S.S.: A new view of automatic relevance determination. In: NIPS. pp. 1625–1632 (2008)
20. Marlin, B.M., Khan, M.E., Murphy, K.P.: Piecewise bounds for estimating bernoulli-logistic latent gaussian models. In: ICML. pp. 633–640 (2011)
21. Bishop, C.M.: Neural networks for pattern recognition. Oxford press (1995)
22. Du Buf, H., Bayer, M., Droop, S., Head, R., Juggins, S., Fischer, S., Bunke, H., Wilkinson, M., Roerdink, J., Pech-Pacheco, J., et al.: Diatom identification: a double challenge called adiac. In: ICIAP. pp. 734–739. IEEE (1999)
23. Tipping, M.E., Bishop, C.M.: Mixtures of probabilistic principal component analyzers. NECO 11(2), 443–482 (1999)
24. Taha, A.A., Hanbury, A.: Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. BMC-MI 15(1), 29 (2015)