Shape Regression Machine

S. Kevin Zhou and Dorin Comaniciu

Integrated Data Systems Department,Siemens Corporate Research
755 College Road East, Princeton NJ 08540

Abstract. We present a machine learning approach called *shape regression machine* (SRM) to segmenting in real time an anatomic structure that manifests a deformable shape in a medical image. Traditional shape segmentation methods rely on various assumptions. For instance, the deformable model assumes that edge defines the shape; the Mumford-Shah variational method assumes that the regions inside/outside the (closed) contour are homogenous in intensity; and the active appearance model assumes that shape/appearance variations are linear. In addition, they all need a good initialization. In contrast, SRM poses no such restrictions. It is a two-stage approach that leverages (a) the underlying medical context that defines the anatomic structure and (b) an annotated database that exemplifies the shape and appearance variations of the anatomy. In the first stage, it solves the initialization problem as object detection and derives a regression solution that needs just one scan in principle. In the second stage, it learns a nonlinear regressor that predicts the nonrigid shape from image appearance. We also propose a boosting regression approach that supports real time segmentation. We demonstrate the effectiveness of SRM using experiments on segmenting the left ventricle endocardium from an echocardiogram of an apical four chamber view.

1 Introduction

Deformable shape segmentation is a long-standing challenge in medical imaging. Numerous algorithms have been proposed in the literature to tackle the problem, among which there are three important approaches: the deformable model or snake [1], the Mumford-Shah variational method [2], and the active appearance model (AAM) [3].

The deformable model or snake [1] seeks a parameterized curve $\mathtt{C}(s)$ that minimizes the cost function $\mathcal{E}_{snake}(\mathtt{C})$:

$$\mathcal{E}_{snake}(\mathtt{C}) = \int_0^1 \{-\mu|\nabla \mathtt{I}(\mathtt{C}(s))|^2 + w_1(s)|\mathtt{C}'(s)|^2 + w_2(s)|\mathtt{C}''(s)|^2\} ds, \quad (1)$$

where μ controls the magnitude of the potential, ∇ is the gradient operator, $\mathtt{I}$ is the image, $w_1(s)$ controls the tension of the curve, and $w_2(s)$ controls the rigidity of the curve. The implicit assumption of the snake model is that edge defines the curve due to the use of the gradient operator.

N. Karssemeijer and B. Lelieveldt (Eds.): IPMI 2007, LNCS 4584, pp. 13–25, 2007.

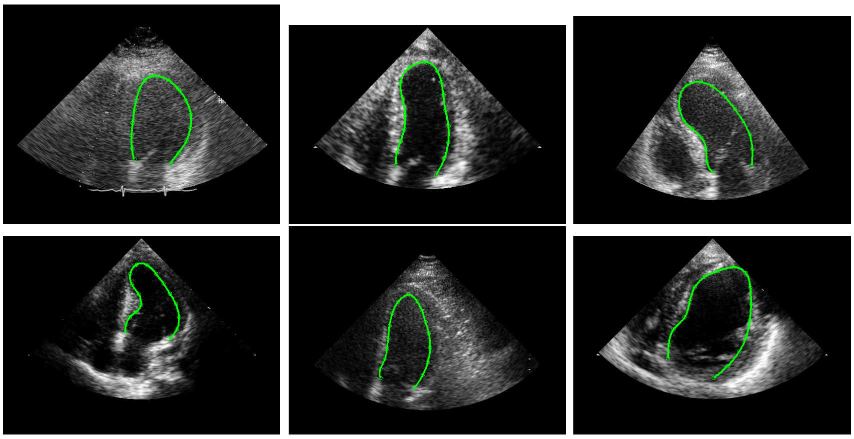


Fig. 1. Examples of A4C echocardiogram. The expert annotation of the LV endocardium is marked by the green line. The shape is represented by 17 landmarks and the cubic spline is used for intepolation.

In the Mumford-Shah variational method [2], the minimal partition problem is mostly studied, where a curve $\mathtt{C}$ is sought to minimize the cost function $\mathcal{E}_{ms}(\mathtt{C})$:

$$\mathcal{E}_{ms}(\mathtt{C}) = \int_{\Omega_i} |\mathtt{I}(x,y) - u_i|^2 dxdy + \int_{\Omega_o} |\mathtt{I}(x,y) - u_o|^2 dxdy + \mu\mathcal{L}(\mathtt{C}), \quad (2)$$

where Ω_i and Ω_o denote the inside and outside regions, respectively, with respect to the curve $\mathtt{C}$, u_i and u_o are piecewise constants for the two regions, and $\mathcal{L}(\mathtt{C})$ is the length of the curve. The region homogeneity is assumed here.

The AAM [3] jointly characterizes the appearance $\mathtt{I}$ and shape $\mathtt{C}$ using a linear generative model:

$$\mathtt{C} = \bar{\mathtt{C}} + \mathtt{Q}_c\mathtt{a}; \quad \mathtt{I} = \bar{\mathtt{I}} + \mathtt{Q}_i\mathtt{a}, \quad (3)$$

where $\bar{\mathtt{C}}$ is the mean shape, $\bar{\mathtt{I}}$ the mean appearance in a normalized patch, and $\mathtt{a}$ is the blending coefficient vector shared by both the shape and appearance. The model parameter $\mathtt{a}$, along with a similarity transformation parameter, is found by fitting the AAM to the observed image using the mean square error criterion.

However, the above assumptions are easily violated in practice. Consider the problem of segmenting the left ventricle (LV) endocardium from an echocardiogram of an apical four chamber (A4C) view. The echocardiogram is an ultrasound image of human heart and the A4C view is a canonical view in which all four heart chambers are visible. Fig. 1 presents several A4C examples that manifest the following facts: (i) The LV endocardium is not defined by the edge. For example, it cuts the papillary muscle attached to the LV; (ii) The region homogeneity is severely violated due to ultrasound imaging artifacts and signal dropouts; and (iii) The shape and appearance variations are hardly linear due to differences in instrument, patient, and sonograher, respiratory interferences, unnecessary probe movements, etc. Furthermore, the above three methods need a good initialization and different initializations might yield very different results due to the attraction of local minima.

In this paper, we present a machine learning approach called shape regression machine (SRM), which poses none of the above restrictions. It deals with deformable contour not necessarily supported by the edge, allows region inhomogeneity, and utilizes nonlinear models to characterize shape and appearance in a discriminative manner. In addition, it is fully automatic with no manual initialization and runs in real time. SRM is mostly appropriate for segmenting an anatomical structure. The core of SRM is to effectively leverage the underlying *structural context* present in medical images and, using regression, to extract knowledge from an *annotated database* that exemplifies the shape and appearance variations. Section 2 depicts the principle of the SRM approach and section 3 elaborates an image-based boosting regression method that underpins SRM. Section 4 presents the experimental results of segmenting the LV endocardium from the A4C echocardiogram.

2 Shape Regression Machine

The shape $\mathtt{C}$ is represented by two parts: rigid and nonrigid. The rigid transformation is parameterized by θ and the nonrigid part by $\mathtt{S}$. If the rigid similarity transformation is used, then the above shape representation reduces to Kendall's interpretation. To rigidly align the LV shape in the A4C echocardiogram more accurately, we use a 5D-parameterization $\theta = (t_x, t_y, \log(s_x), \log(s_y), \alpha)$, with (t_x, t_y) for translation, α for orientation, and (s_x, s_y) for scale (or size) in both x- and y-directions. Due to the multiplicative nature of the scale parameter, we take the logarithm operator to convert it to additive. Fig. 2(a) illustrates the meaning of the five parameters.

SRM is a two-stage approach. It first solves the rigid transformation θ as object detection and then infers the nonrigid part $\mathtt{S}$, both using the machine learning technique of regression.

2.1 Regression-Based Object Detection

A promising approach to medical anatomy detection is to use the classifier-based object detection approach like [4]: It first trains a binary classifier, discriminating the anatomic structure of interest from the background, and then exhaustively scans the query image for anatomy targets. In [4], the so-called integral image is proposed to enable real time evaluation of the classifier when applied for searching the translation parameter exhaustively and the scale parameter sparsely. No orientation is scanned. However, the medical anatomy such as LV often manifests arbitrary orientation and scale. To give an accurate account of orientation and scale, which is required for subsequent tasks like LV endocardial wall segmentation, the detection speed is sacrificed if a dense set of orientations and scales is tested. In general, the computational complexity of the classifier-based approach linearly depends on the image size (for the translation parameter), and the number of tested orientations and scales. Also, multiple integral images according to different rotations need to be computed. Therefore, the bottleneck of the classifier-based detection approach lies in its exhaustive scanning native. To avoid exhaustive scanning, we propose a regression-based detection approach. By leveraging the anatomical structure that manifests regularization and context in geometry and appearance, we formulate a novel regression task that, in theory, necessitates *only one scan.* Also, we compute *only one integral image.*

Basic idea. Fig. 2(b) demonstrates the basic idea of the regression-based medical anatomy detection. For illustrative purpose only, we address only the translation parameter θ as in Fig. 2(b). In other words, we are only interested in finding the center position $\theta_0 = (t_{x,0}, t_{y,0})$ of the LV in an A4C echocardiogram, assuming that the orientation of the LV is upright and the scale/size of the LV is fixed. It is straightforward to extend the 2D case to the 5D-parameterization.

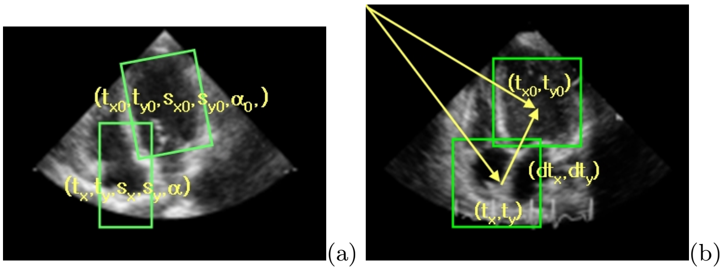

$(t_{x0},t_{y0},s_{x0},s_{y0},\alpha_0,)$
(t_x,t_y,s_x,s_y,α)
(a)
(t_{x0},t_{y0})
(dt_x,dt_y)
(t_x,t_y)
(b)

Fig. 2. (a) The regression setting of a 5D parameter space: (t_x, t_y) is the LV center, (s_x, s_y) the LV size, and α the LV angle. (b) A graphical illustration of regression-based medical anatomy detection based on a 2D translation parameterization.

Suppose that, during running time, we confront an image patch $\mathtt{I}(\theta)$ centered at position $\theta = (t_x, t_y)$. If there exists an oracle $\mathcal{F}_1$ that does the following: given an image patch $\mathtt{I}(\theta)$, it tells the difference vector $d\theta$ between the current position θ and the target position $\theta_0 = (t_{x,0}, t_{y,0})$, i.e., $d\theta = \theta_0 - \theta$, then we achieve the detection using *just one* scan. In other words, through the oracle that defines a mapping $\mathcal{F}_1 : \mathtt{I} \longrightarrow d\theta$, the ground truth position $\hat{\theta}_0$ is estimated as follows.

$$d\theta = \mathcal{F}_1(\mathtt{I}(\theta)), \quad \hat{\theta}_0 = \theta + d\theta = \theta + \mathcal{F}_1(\mathtt{I}(\theta)). \tag{4}$$

Learning the function $\mathcal{F}_1(\mathtt{I}(\theta))$ is referred to as *regression* in machine learning.

Does such an oracle $\mathcal{F}_1$ exist? Since the anatomic structure of interest is tied with human body atlas, there is a known number of objects appearing within geometric and appearance contexts. Often only one object is available. For example, in the A4C echocardiogram, there is only one target LV available and its relation with respect to other structures such as left atrium, right ventricle and right atrium is geometrically fixed (that is why they are called left/right ventricle/atrium). Also there exists a strong correlation among their appearances. By knowing where the LA, RV, or RA is, we can predict the LV position quite accurately. In principle, by knowing where we are (i.e., knowing θ) and then looking up the map/atlas that tells the difference to the target (i.e., telling $d\theta$ through the oracle), we can reach the target instantaneously in a virtual world.

Medical atlas is widely used in the literature [5,6]. However, the methods in [5,6] use the atlas as an *explicit* source of prior knowledge about the location, size, and shape of the anatomic structures and deform it to match the image content for registration, segmentation, tracking, etc. In this paper, we take an *implicit* approach, that is, embedding the atlas in a learning framework. After learning, the atlas knowledge is fully absorbed and the atlas is no longed kept.

How to learn the oracle $\mathcal{F}_1$? We leverage machine learning techniques, based on an annotated database. As in Fig. 3, we first collect from the database input-output pairs (as many as possible) as training data. By varying the location, we crop out different local image patches while recording their corresponding difference vectors. Similarly, for the 5D parameterization, we can extract the training data. We now confront a multiple regression setting with a multidimensional output, which is not well addressed in the machine learning literature. In this paper, we propose the image-based boosting regression (IBR) algorithm to fulfill the learning task.

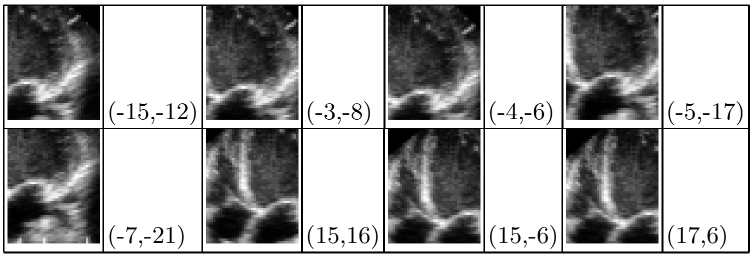


Fig. 3. Training image examples (generated based on the image in Fig. 2(b)): image I and its associated rigid transformation parameter $d\theta = (dx, dy)$

Detection algorithm. In theory, only one scan is needed to find the target; in practice, we conduct a sparse set of random scans and then estimate the parameter using fusion. Suppose that in total M random samples are scanned at positions $\{\theta^{<1>}, \theta^{<2>}, \dots, \theta^{<M>}\}$. For each $\theta^{<m>}$, we invoke the regressor to predict the difference parameter $d\theta^{<m>}$ and, subsequently, the target parameter $\theta_0^{<m>}$ as follows:

$$d\theta^{<m>} = \mathcal{F}_1(\mathtt{I}(\theta^{<m>})), \quad \theta_0^{<m>} = \theta^{<m>} + d\theta^{<m>}, \quad m = 1, 2, \dots, M. \qquad (5)$$

We also learn a binary classifier (or detector) $\mathcal{D}$ that separates the object from the background and use its posterior probability $p_d(\mathtt{I})$ of being positive as a confidence scorer. After finding the m^{th} prediction $\theta_0^{<m>}$, we apply the detector $\mathcal{D}$ to the image patch $\mathtt{I}(\theta_0^{<m>})$. If the detector $\mathcal{D}$ fails, we discard the m^{th} sample; otherwise, we keep the confidence score $p_d^{<m>}$. This way, we have a weighted set

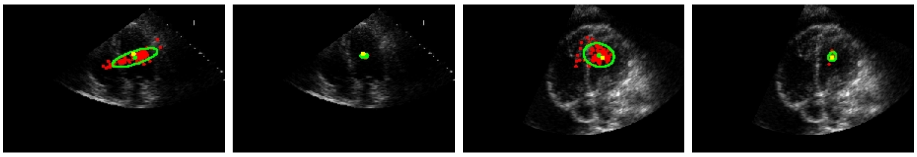


Fig. 4. The odd-indexed images show the 100 predicted target outputs (red) and the even-indexed images show only the predicted target outputs (red) passing the detector. The green point is the final estimate of the target position, the green curve is the 95% confidence curve, and the yellow point indicates the ground truth position. Note that the region bounded the 95% confidence curve on the even-indexed images is significantly smaller than that on the odd-indexed images.

$\{(\theta_0^{<j>}, p_d^{<j>}); j = 1, 2, \ldots, J\}$ (note that $J \leq M$ as samples might be dropped), based on which we calculate the weighted mean as the final estimate $\hat{\theta}_0$

$$\hat{\theta}_0 = \{ \sum_{j=1:J} p_d^{<j>} \theta_0^{<j>} \} / \{ \sum_{j=1:J} p_d^{<j>} \}. \qquad (6)$$

In practice, we stop scanning when $J \geq J_{valid}$ in order to further save computation. If there is no sample $\theta_0^{<m>}$ passing $\mathcal{D}$, then we use the unweighted mean of $\theta_0^{<m>}$ as the final estimate.

Combining the regressor and binary detector yields an effective tool for medical anatomy detection; empirical evidence tells that, when compared with the method using only the regressor, it needs only a smaller number of scans to reach a better performance. Fig. 4 demonstrates the intuition behind this improvement using the 2-D translational case. Two example images are shown along with their $M = 100$ predicted target positions (the red points). The majority of the prediction is close to the ground truth position (the yellow point) although there are outliers. Fig. 4 also shows the predicted points passing the detector: All the outliers are eliminated, thereby significantly improving the precision of the estimate as evidenced by the smaller region bounded by the 95% confidence curve.

2.2 Regression-Based Nonrigid Shape Inference

After the first stage that finds the bounding box (parameterized by θ) to contain the object, we have the object rigidly aligned. In the second stage, we are interested in inferring the nonrigid part $\mathtt{S}$. In this paper, we assume that $\mathtt{S}$ consists of N landmark points, i.e., $\mathtt{S} = [x_1, y_1, \ldots, x_N, y_N]^{\mathsf{T}}$. Other shape representations can be used with no difficulty.

Basic idea. We formulate the nonrigid shape inference again as a regression problem. In other words, we seek an oracle $\mathcal{F}_2$ that tells the shape $\mathtt{S}$ based on the image patch $\mathtt{I}$ that is known to contain the object.

$$\mathtt{S} = \mathcal{F}_2(\mathtt{I}). \qquad (7)$$

Does such an oracle $\mathcal{F}_2$ exist? Because we deal with one particular anatomic structure (say LV), it is obvious that a regularity exists in terms of its appearance

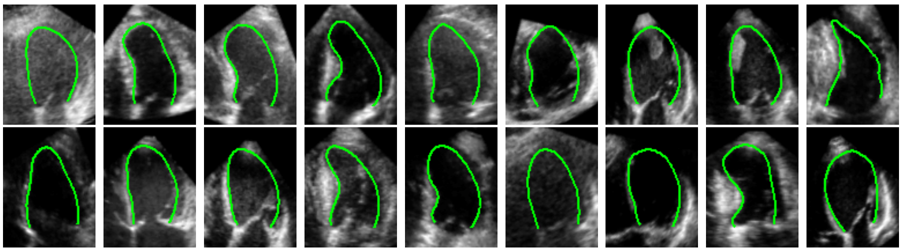


Fig. 5. Training image examples: image $\mathtt{I}$ and its associated nonrigid shape $\mathtt{S}$. The first six images correspond to those in Fig. 1. The image size is 104 by 80.

and shape although the variations in them can be quite significant. Fig. 5 displays several images with corresponding shapes that are rigidly aligned to the mean shape. As mentioned earlier, a linear modeling of the appearance and shape is insufficient. One goal of the paper is to provide a nonlinear modeling of the shape and appearance.

How to learn the oracle $\mathcal{F}_2$? Given an annotated database, we extract corresponding pairs of (already rigidly aligned) shape and appearance as in Fig. 5. We also slightly perturb the rigid parameter to accommodate imperfect localization derived from the first stage. We now again confront a multiple regression setting with a multidimensional output, except that this time the output cardinality is even higher.

Inference algorithm. To improve robustness, we slightly perturb the bounding box[1] to generate K random samples $\{\mathtt{I}^{<1>}, \mathtt{I}^{<2>}, \ldots, \mathtt{I}^{<K>}\}$ and apply the regressor to obtain shape estimates $\{\mathtt{S}^{<1>}, \mathtt{S}^{<2>}, \ldots, \mathtt{S}^{<K>}\}$, where $\mathtt{S}^{<k>} = \mathcal{F}_2(\mathtt{I}^{<k>})$. We also build a nonparametric density $p_s(\mathtt{S})$ based on the prior shape examples and use it as a confidence scorer. Finally, we output the weighted mean as the final estimate $\hat{\mathtt{S}}$ for the shape parameter (we empirically choose $K = 10$):

$$\hat{\mathtt{S}} = \{ \sum_{k=1:K} p_s^{<k>} \mathtt{S}^{<k>} \} / \{ \sum_{k=1:K} p_s^{<k>} \}. \tag{8}$$

3 Image-Based Boosting Regression

The underpinning of the above two stages of SRM is a regression procedure that takes an image as input and outputs a multidimensional variable. In this section, we invoke the influential boosting framework [7,8] to derive a novel regression algorithm called image-based boosting regression (IBR).

We denote a scalar by a, a column vector by $\mathtt{a}$, and a matrix by $\mathtt{A}$. We also denote the input by $\mathtt{x} \in \mathcal{R}^d$, the output by $\mathtt{y}(\mathtt{x}) \in \mathcal{R}^q$, the regression function by $\mathtt{g}(\mathtt{x}) : \mathcal{R}^d \rightarrow \mathcal{R}^q$ and the training data points by $\{(\mathtt{x}_n, \mathtt{y}_n); n = 1, 2, ..., N\}$.

[1] The perturbation is limited to translation and scaling as they share one integral image. There is no perturbation in rotation.

Further, we denote $\mathtt{x}^\top\mathtt{x} = \|\mathtt{x}\|^2$ and $tr(\mathtt{X}^\top\mathtt{X}) = \|\mathtt{X}\|^2$. In SRM, $\mathtt{x}$ is the image $\mathtt{I}$, $\mathtt{y}$ is the difference vector $d\theta$ in the first stage and the nonrigid shape parameter $\mathtt{S}$ in the second stage, and the regression function $\mathtt{g}(\mathtt{x}) = \mathcal{F}(\mathtt{I})$ is the oracle.

IBR minimizes the following cost function, which combines a regression output fidelity term and a regularization term:

$$J(\mathtt{g}) = \sum_{n=1:N} \{\|\mathtt{y}(\mathtt{x}_n) - \mathtt{g}(\mathtt{x}_n)\|^2\} + \lambda R(\mathtt{g}), \tag{9}$$

where λ is a *regularization coefficient* and $R(g)$ is the regularization term that will be subsequently defined. As in any boosting procedure [7,8], IBR assumes that the regression output function $\mathtt{g}(\mathtt{x})$ takes an additive form:

$$\mathtt{g}_t(\mathtt{x}) = \mathtt{g}_{t-1}(\mathtt{x}) + \mathtt{h}_t(\mathtt{x}) = \sum_{i=1:t} \mathtt{h}_i(\mathtt{x}), \tag{10}$$

where each $\mathtt{h}_i(\mathtt{x}) : \mathcal{R}^d \rightarrow \mathcal{R}^q$ is a weak learner (or weak function) residing in a *dictionary* set $\mathcal{H}$, and $\mathtt{g}(\mathtt{x})$ is a strong learner (or strong function).

Boosting [7,8] is an iterative algorithm that leverages the additive nature of $\mathtt{g}(\mathtt{x})$: At iteration t, one more weak function $\mathtt{h}_t(\mathtt{x})$ is added to the target function $\mathtt{g}(\mathtt{x})$ to maximally reduce the cost function. Because we associate each weak function with visual features (as shown next), boosting operates as a *feature selector* that singles out relevant features to the regression task.

Weak function. We use a bank of over-complete features to represent the image $\mathtt{x}$. In particular, we use the Haar-like local rectangle features [4], whose rapid evaluation is enabled by the use of integral image. As shown in [4], (i) it is easy to construct numerous local rectangle features and (ii) the local rectangle feature, whose response is normalized by the standard deviation of the image patch, is relatively robust to appearance variation. Each local rectangle feature $f(\mathtt{x};\mu)$ has its own attribute μ, namely feature type and window position/size.

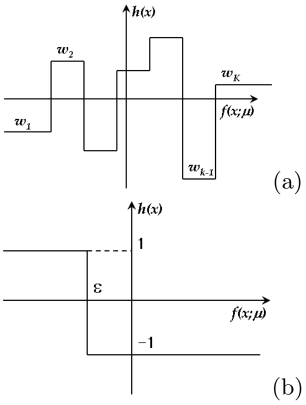
Based on the local rectangle features, we construct one-dimensional (1D) regression stumps as primitives of the dictionary set $\mathcal{H}$. A regression stump $h(\mathtt{x};\mu)$, illustrated in Fig. 6(a), is defined as

$$h(\mathtt{x};\mu) = \sum_{k=1:K} w_k \; [f(\mathtt{x};\mu) \in R_k] = \mathtt{e}(\mathtt{x};\mu)^\top \mathtt{w}, \tag{11}$$

where [.] is an indicator function and $\{R_k;\ k = 1, 2 \ldots, K\}$ are K evenly spaced intervals (except that R_1 and R_K go to ∞). The interval boundary points are empirically determined. We first find the minimum and maximum responses for the feature and then uniformly divide them. In (11), all the weights w_k are compactly encoded by a vector $\mathtt{w}_{K\times 1} = [w_1, w_2, \ldots, w_K]^\top$ and the vector $\mathtt{e}(\mathtt{x};\mu)$ is some column of the identity matrix: only one element is 1 and others are 0.

A weak function is constructed as a q-dimensional (q-D) regression stump $\mathtt{h}(\mathtt{x})_{q\times 1}$ that stacks q different 1D regression stumps, i.e.,

$$\mathtt{h}(\mathtt{x};\mu_1,\ldots,\mu_q) = [h_1(\mathtt{x};\mu_1),...,h_q(\mathtt{x};\mu_q)]^\top = [\mathtt{e}_1(\mathtt{x};\mu_1)^\top\mathtt{w}_1,...,\mathtt{e}_q(\mathtt{x};\mu_q)^\top\mathtt{w}_q]^\top, \tag{12}$$



```
1. Initialization $t = 0$.
   (a) Set $\lambda$ (the regularization coefficient) and $\eta$ (the shrinkage factor).
   (b) Set the values related to the stopping criteria: $T_{max}$ (the maximum number of iterations) and $J_{min}$ (the minimum cost function).
   (c) Set initial values for $t = 0$: $\mathtt{g}_0(\mathtt{x}) = 0$ and $\mathtt{r}_0(x) = \mathtt{y}(\mathtt{x})$.
2. Iteration $t = 1, \ldots, T_{max}$
   (a) Find the optimal $\hat{\mathtt{h}}_t$ that solves (13).
   (b) Form the new function $\mathtt{g}_t(\mathtt{x}) = \mathtt{g}_{t-1}(\mathtt{x}) + \eta\hat{\mathtt{h}}_t(\mathtt{x})$.
   (c) Evaluate the approximation error $\mathtt{r}_t(\mathtt{x}) = \mathtt{y}(\mathtt{x}) - \mathtt{g}_t(\mathtt{x})$ and the cost function $J(\mathtt{g}_t)$.
   (d) Check convergence, e.g., see if $J(\mathtt{g}_t) < J_{min}$.
```
(c)

Fig. 6. (a) Regression stump. (b) Binary decision stump. The regression stump carries more representational power than the decision stump. (c) The proposed image-based boosting regression (IBR) algorithm.

where $\mathtt{w}_j$ is the weight vector for the j^{th} regression stump $h_j(\mathtt{x}; \mu_j)$. We further encode the weights belonging to all regression stumps into a *weight matrix* $\mathtt{W}_{K\times q} = [\mathtt{w}_1, \mathtt{w}_2, \ldots, \mathtt{w}_q]$. A binary decision stump is used in [4]. Fig. 6(a,b) compares the regression and binary decision stumps.

Boosting ridge regression. The model complexity of the regression output function $\mathtt{g}_t(\mathtt{x}) = \sum_{i=1:t} \mathtt{h}_i(\mathtt{x})$ now depends on its weight matrices $\{\mathtt{W}_i, i = 1, \ldots, t\}$. We incorporate the ridge regression principle [9] (also known as Tikhonov regularization) into a boosting framework to penalize overly complex models. Because boosting regression proceeds iteratively, at the t^{th} boosting iteration, we set up the following ridge regression task that only involves the weight matrix $\mathtt{W}_t$:

$$\arg\min_{\mathtt{W}_t}\{J_t(\mathtt{g}) = \sum_{n=1:N} \{\|\mathtt{r}_t(\mathtt{x}_n) - \mathtt{h}_t(\mathtt{x}_n)\|^2\} + \lambda\|\mathtt{W}_t\|^2\}, \tag{13}$$

where $\mathtt{r}_t(\mathtt{x}_n) = \mathtt{y}(\mathtt{x}_n) - \mathtt{g}_{t-1}(\mathtt{x}_n)$ is the residual.

As the weight vectors $\{\mathtt{w}_1, \mathtt{w}_2, \ldots, \mathtt{w}_q\}$ in the matrix $\mathtt{W}_t$ are associated with q different local rectangle features, the optimization in (13) implies two subtasks:

1. Given a set of q features with attributes $\mu_1, \ldots, \mu_q$, respectively, find the optimal matrix $\hat{\mathtt{W}}_t(\mu_1, \ldots, \mu_q)$ and its minimum cost $\hat{J}_t(\mu_1, \ldots, \mu_q)$;
2. Find the optimal set of q features with respective attributes $\hat{\mu}_1, \ldots, \hat{\mu}_q$ that minimizes the cost $\hat{J}_t(\mu_1, \ldots, \mu_q)$. This corresponds to feature selection.

However, to transform the above optimization into an efficient implementation, there is a computational bottleneck: The second subtask necessitates a greedy feature selection scheme, which is too expensive to evaluate given a large local rectangle feature pool. In practice, approximate non-greedy solutions [10] can be derived to speedup the feature selection process; however, this is beyond the scope of the paper. Finally, IBR invokes shrinkage [9] to derive a smooth output: $\mathtt{g}_t(\mathtt{x}) = \mathtt{g}_{t-1}(\mathtt{x}) + \eta \mathtt{h}_t(\mathtt{x})$. Fig. 6(c) summarizes the IBR algorithm.

4 Experimental Results and Discussions

We applied the SRM approach to segmenting the LV endocardium from 2D echocardiograms. We had in total 527 A4C sequences. Though we had video sequences, we focused on detecting the LV at the end of diastole (ED) frame, when the LV dilates to its maximum. We randomly selected 450 ED frames for training and used the remaining 77 for testing.

4.1 Rigid Object Detection

In this experiment, we tested the first stage of SRM to detect the LV using the 5-D parameterization. Figure 1 shows six ED images with the unaligned LV present. The range of the five parameters is empirically found as: $t_x \sim [43, 118]$, $t_y \sim [24, 70]$, $s_x \sim [26, 86]$, $s_y \sim [37, 92]$ and $\alpha \sim [-25, 35]$. We scanned the image following the above range. The average image size is 111×151.

There are several tuning parameters in the IBR algorithm. For the number of threshold levels K of a weak function, the regularization coefficient λ and the shrinkage coefficient η, we empirically tested different combinations and decided to use the following: $K = 64$, $\lambda = 0.1/K$, and $\eta = 0.1$. We trained the regressor based on 450 randomly selected ED frames, each yielding 30 image patches; in total we had 13,500 training data. It takes more than two days to train the regressor (on a high-end workstation with four Xeon 3GHz CPUs and 3GB RAM), which consists of 10,000 local rectangle features or 200 weak functions. Training the detector $\mathcal{D}$ is not straightforward because here the image rotation is involved. To avoid computing integral images for all rotations, we followed [11] to train the detector, which is able to simultaneously classify the object as well as infer its rotation yet using only one integral image.

We implemented three scanning methods: "IBR", "IBR+Det", and "Det". The "IBR" means that we randomly scanned the image within the prior range using the learned IBR function and used the unweighted average as the final estimate of the target position. The "IBR+Det" means that we further equipped the "IBR" method with the trained detector and used (6) as the final estimate. We also set $J_{valid} = 10$ to enable early exit when scanning. The "Det" means that we exhaustively scanned the image within the same range using the detector and used the parameter that maximizes the detector response as the final estimate. For the "Det" method, we exhaustively scanned the image every 4 pixels in both translations and every 4 pixels in both scales.

Table 1(a) compares the three scanning methods[2]. The error in scale is measured as $s_{detected}/s_{groundtruth} - 1$. Because we did not observe significant performance difference between training and testing, we pooled them together and

[2] To count the number of effective scans in Table 1, we excluded those scans if their associated image patches have less than 40% of their pixels inside the known fan.

Table 1. (a) Detection performance comparison of the three methods for the 5-parameter case. (b) Segmentation performance comparison of four regression methods.

Method	IBR	IBR+Det	Det
# of features	10000	10000+1201	1201
median err. in t_x (pixels)	0.32 ± 3.13	0.65 ± 2.07	1.69 ± 3.40
median err. in t_y (pixels)	0.67 ± 2.40	1.25 ± 1.95	0.84 ± 3.73
median err. in s_x	0.02 ± 0.12	0.04 ± 0.12	0.05 ± 0.17
median err. in s_y	0.01 ± 0.08	0.02 ± 0.08	0.04 ± 0.15
median err. in α (degree)	-1.76 ± 7.17	-0.98 ± 6.39	0.22 ± 6.74
# of eff. scans	200	38	29383
avg. speed (ms)	704	118	6300

(a)

Method	SRM	KRR	NPR	AAM
25% seg. err. (pixels)	1.778	1.695	2.013	2.323
median seg. err. (pixels)	2.207	2.372	2.482	2.734
75% seg. err. (pixels)	2.753	3.347	3.101	4.002
avg. speed (ms)	≤ 1	692	865	30

(b)

jointly reported the results. The speed was recorded on a laptop with a Pentium 2.1GHz CPU and 2GB RAM. The "IBR+Det" achieves appealing detection performance while running the fastest. It runs about 7 times faster than the "IBR" method and more than 50 times faster than the "Det" method, while yielding comparable performance to the "IBR" in terms of bias and improving the localization precision. The slowest "Det" method does not always yield the best performance in terms of either bias or variance because it does not exhaust all possible configurations. Fig. 7(a) shows example images with estimated and ground truth boxes overlaid.

4.2 Nonrigid Shape Inference

In this experiment, we invoked the complete SRM approach to automatically delineate the LV endocardium. The above "IBR+Det" algorithm was first used to locate the LV and then the second stage of SRM was applied. The shape $\mathtt{S}$ is parameterized by 17 landmark points and PCA was used to reduced the shape dimensionality from 34 to 20. Through random perturbations, we generated 6,750 training data points (one data point is a pair of image and shape) based on 450 ED frames and trained an IBR model consisting of 20,000 local rectangle features or 1,000 weak functions.

For comparison, we implemented three other regression methods: "KRR", "NPR", and "AAM" where kernel rigid regression (KRR) and nonparametric kernel regression (NPR) are two off-the-shelf nonlinear regression methods [9], and AAM is from [3]. In AAM, the appearance and shape are assumed to be jointly Gaussian, which amounts to multiple linear regression [9]. The number of principal components was chosen to keep 95% of the energy in AAM. When comparing different nonrigid shape regressors, we fixed the detection part.

To quantify the shape segmentation performance, we measured the average pixel error for the landmark points: $\sqrt{||\mathtt{C}_1 - \mathtt{C}_2||^2/34}$. We did this measurement on the aligned domain of size 104 by 80 to overcome the difference in physical

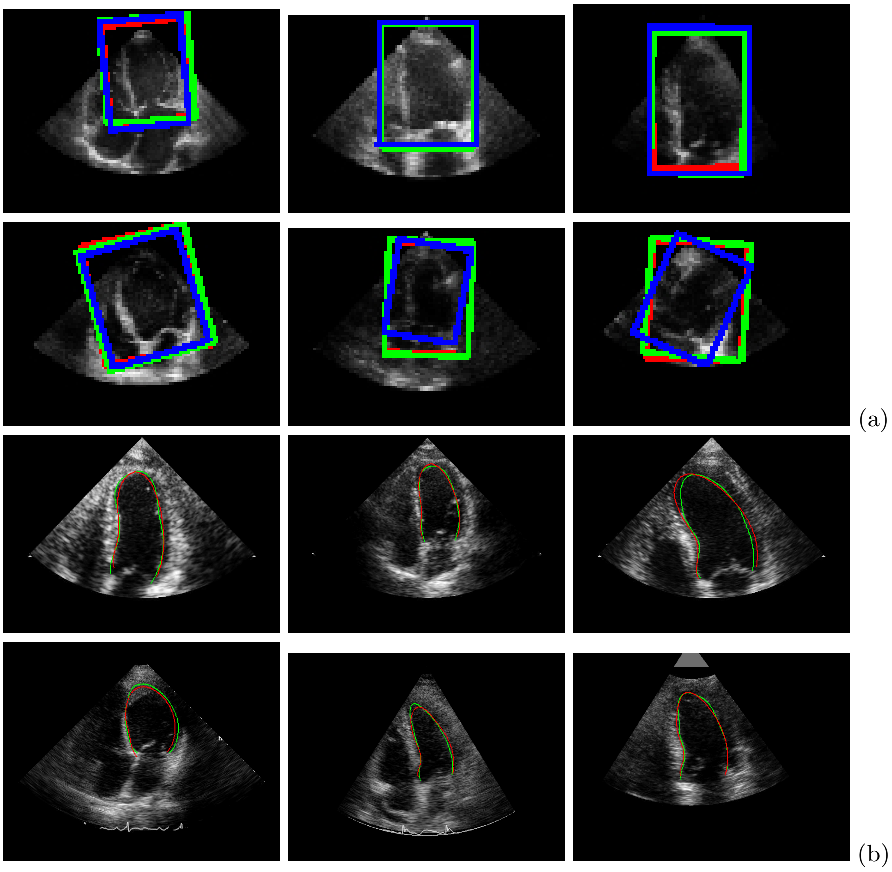


Fig. 7. (a) The estimated LV box versus the ground truth. The red box is from the "IBR" method, the green is from the "IBR+Det" method, and the blue is the ground truth. (b) The inferred LV endocardium versus the ground truth. The red contour is from the SRM approach and the green is the ground truth.

units of difference images. Table 1(b) shows the error statistics and computational time (only for the regression part though). We collected the error statistics for all testing images and reported their 25% percentile, median, and 75% percentile. From Table 1(b), we observe that the proposed SRM approach achieves favorable contour localization performance over other methods while running significantly faster. The AAM method that uses linear models performs the worst, implying the need for nonlinear modeling of the appearance and shape. The KRR and NPR methods are slow because they require comparing the query image with the whole database, while the IBR absorbs the database knowledge into the weak functions whose rapid evaluation is guaranteed by using the integral image. In sum, *it takes less than 120ms on the average to automatically localize the LV endocardium in an A4C echocardiogram with a better accuracy.* Fig. 7(b) visualizes the ground truth and predicted contours.

5 Conclusion

We have presented a machine learning approach called shape regression machine for fast medical anatomy detection and segmentation. SRM effectively utilizes the structural context in medical images with annotations to eliminate unfavorable restrictions posed by conventional deformable shape segmentation methods. In particular, the detection solution in SRM replaces the exhaustive scanning of the query image required by the classifier-based detector by a sparse scanning and reaches improved detection accuracy with significantly less computation and no need for image rotation. In terms of shape inference, the IBR solution in SRM outperforms other regression methods such as kernel ridge regression, nonparametric kernel regression, and active appearance model. In the future, we will apply the SRM approach to other medical applications such as organ segmentation from a full body 3D CT scan. We will also address the scalability and trainability issues related to learning the regression function.

References

1. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. Int. J. Computer Vision 1, 321–331 (1988)
2. Mumford, D., Shah, J.: Optimal approximations by piecewise smooth functions and associated variational problems. Comm. Pure Appl. Math. 42, 577–685 (1989)
3. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Trans. Pattern Anal. Machine Intell. 23, 681–685 (2001)
4. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proc. CVPR (2001)
5. Cootes, T., Beeston, C., Edwards, G., Taylor, C.: A unified framework for atlas matching using active appearance models. In: IPMI (1999)
6. Mazziotta, J., Toga, A., Evans, A., Lancaster, J., Fox, P.: A probabilistic atlas of the human brain: Theory and rational for its development. Neuroimage 2, 89–101 (1995)
7. Freund, Y., Schapire, R.: A decision-theoretic generalization of online leaning and an application to boosting. J. Computer and System Sciences 55, p. 119
8. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting. Ann. Statist. 28, 337–407 (2000)
9. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. Springer, Heidelberg (2001)
10. Zhou, S., Georgescu, B., Zhou, X., Comaniciu, D.: Image-based regression using boosting method. In: Proc. ICCV (2005)
11. Zhang, J., Zhou, S., Comaniciu, D.: Joint real-time object detection and pose estimation using probabilistic boosting network. In: Proc. CVPR (2007)