WERITAS - Weighted Ensemble of Regional Image Textures for ASM Segmentation

Robert Toth[a], Scott Doyle[a], Mark Rosen[b], Arjun Kalyanpur[c], Sona Pungavkar[d], B. Nicolas Bloch[e], Elisabeth Genega[e], Neil Rofsky[e], Robert Lenkinski[e], Anant Madabhushi[a]

[a]Rutgers, The State University of New Jersey, NJ, USA. [b]University of Pennsylvania, Philadelphia, PA, USA. [c]Teleradiology Solutions, Bangalore, India. [d]Dr. Balabhai Nanavati Hospital, Mumbai, India. [e] Department of Radiology, Beth Israel Deaconess Medical Center, MA, USA.

ABSTRACT

In this paper we present WERITAS, which is based in part on the traditional Active Shape Model (ASM) segmentation system. WERITAS generates multiple statistical texture features, and finds the optimal weighted average of those texture features by maximizing the correlation between the Euclidean distance to the ground truth and the Mahalanobis distance to the training data. The weighted average is used a multi-resolution segmentation system to more accurately detect the object border. A rigorous evaluation was performed on over 200 clinical images comprising of prostate images and breast images from 1.5 Tesla and 3 Tesla MRI machines via 6 distinct metrics. WERITAS was tested against a traditional multi-resolution ASM in addition to an ASM system which uses a plethora of random features to determine if the selection of features is improving the results rather than simply the use of multiple features. The results indicate that WERITAS outperforms all other methods to a high degree of statistical significance. For 1.5T prostate MRI images, the overlap from WERITAS is 83%, the overlap from the random features is 81%, and the overlap from the traditional ASM is only 66%. In addition, using 3T prostate MRI images, the overlap from WERITAS is 77%, the overlap from the random features is 54%, and the overlap from the traditional ASM is 59%, suggesting the usefulness of WERITAS. The only metrics in which WERITAS was outperformed did not hold any degree of statistical significance. WERITAS is a robust, efficient, and accurate segmentation system with a wide range of applications.

Keywords: Active shape models (ASMs), *in vivo* MRI, Segmentation, Feature Selection, Texture Features

1. INTRODUCTION

Segmentation is the process of defining where an object is in an image. One of the most common segmentation systems is the Active Shape Model (ASM) segmentation scheme.[1] The ASM model segments an object in a 2 step fashion: detect the object border, and update the shape to fit the detected border. Our system focuses on improving the border detection part of the ASM. Normally, landmarks are placed on the border by experts on a series of training images. Then, pixels in the normal direction are sampled and a statistical appearance model is created. When searching for the border in a test image, the Mahalanobis distance is minimized between the test image's intensity values and the training images' intensity values to detect the border.

While ASM's set the groundwork for a very efficient and accurate segmentation system, there are some inherent limitations. The first is the requirement for a proper initialization. If the system is initialized too far from the ground truth, the system won't be able to converge on the correct object border. Also, the use of the Mahalanobis distance leads to some limitations. First of all, the Mahalanobis distance assumes that the distribution of intensity values is Gaussian, which need not necessarily be the case. ASM's normally find the location with the minimum Mahalanobis distance, and assume that is close to the object border. But outliers, local minima, and intensity artifacts often prevent accurate segmentations. Secondly, image intensities might not necessarily be the optimal texture to use, as intensities are prone to noise and artifacts, which detracts from an accurate segmentation. In addition, with limited training data, the Mahalanobis distance will be undefined if too many pixels are sampled.

Several improvements to traditional ASM's have been proposed. The first is the popular Active Appearance Model (AAM), which creates a global appearance model of the object, and combines that model with the shape information. [2]

Contact Info: robtoth@gmail.com, anantm@rci.rutgers.edu

Medical Imaging 2009: Image Processing, edited by Josien P. W. Pluim, Benoit M. Dawant,
Proc. of SPIE Vol. 7259, 725905 · © 2009 SPIE
CCC code: 1605-7422/09/$18 · doi: 10.1117/12.812473

In addition, the AAM model was improved to be more robust to occlusion and outliers (called Robust AAM), [3] and the ASM model was also independently improved to be more robust to outliers. [4] In addition, casting the entire system in a multi-resolution framework is a simple yet effective method for improving accuracy, [5] which we adopt in our system. If a multi-resolution framework is not desired, methods for initialization such as that presented in [6] can be used to offer a reasonable initialization. A major improvement to the traditional ASM is ASM with Optimal Features (ASMOF), [7] which was shown to offer significant improvements. ASMOF steers clear entirely of using the Mahalanobis distance, and instead creates a classifier as to whether a pixel is considered inside or outside of the object. Then, whichever features best classify pixels are used in the segmentation algorithm. The significant improvements offered by this approach show the usefulness of using features other than just image intensities, although it is unclear whether the improvements come from the features or from using a classifier instead of the Mahalanobis distance. A second segmentation system which builds upon the traditional ASM scheme is Minimal Shape and Intensity Cost Path Segmentation (MISCP). [8] While this system contains many differences to the traditional ASM, we focus on two major improvements. The first improvement is the idea of sampling a neighborhood around each landmark point instead of just pixels along a straight line. This allows more information to be gleaned from the training data.

A second improvement is the idea of using multiple statistical texture features, and during segmentation averaging the Mahalanobis distance values from each feature. We previously proposed a system (MANTRA), which uses multiple statistical texture features, and is driven by using Combined Mutual Information as compared to the Mahalanobis distance. [9] MANTRA showed the advantages of using multiple texture features over simply using intensities. However, there was no intelligent method of feature selection, so while it did offer improvements over the traditional ASM, there is still room for significant increases in accuracy, which we hope to accomplish with WERITAS.

When using multiple features, the problem becomes one of feature selection. While adding features does offer significantly more textural information, some features may detract from the segmentation, and some features might not perform any better than simply using intensities. Our approach is donned WERITAS (from Latin, for ”truth”), and stands for Weighted Ensembled of Regional Image Textures for ASM Segmentation. We build upon the idea presented in MISCP, in which the Mahalanobis distance of multiple features were averaged. But instead of simply taking an average of a large set of features, we find the optimal features to average, and the optimal weights for those features by maximizing the correlation between the Euclidean distance to the ground truth, and the Mahalanobis distance to the training data. The spirit of the system is that with a high correlation, the pixels with a low Mahalanobis distance will consistently be closer to the ground truth. We adopt this idea because ideally the Mahalanobis distance should be minimum at the ground truth, and should increase as the distance to the ground truth increases, so that the ASM can hone in closer on the true object border at each iteration. Our base set of features include 84 1st and 2nd order statistical texture features [10,11] such as mean, median, standard deviation, Haralick entropy, and Sobel gradient, which have been previously shown to be useful in both computer aided diagnosis systems and registration tasks. [10–13] A second improvement to the traditional use of the Mahalanobis distance deals with the limitation that arises with limited training data. Currently, the Mahalanobis distance is undefined if the number of pixels sampled near a landmark point is greater than the number of training images. This happens because the covariance matrix becomes sparse, which makes the inverse impossible to compute. We overcome this limitation by taking the Moore-Penrose pseudo-inverse [14] of the covariance matrix instead of the traditional inverse. The Moore-Penrose pseudo-inverse is a generalized inverse, holding many of the same properties of the traditional inverse, and can be calculated for any matrix, including sparse matrices with determinants of zero. This means that the sampling area will not be limited by the training data. The following outlines the specific novel contributions of the system.

1. The optimal linear combination of features' Mahalanobis distances to increase correlation between Euclidean distance to the ground truth and Mahalanobis distance to the training data.
2. The use of the optimal linear combination of features' Mahalanobis distances in a ASM-based segmentation system.
3. The use of the Moore-Penrose pseudo-inverse in the Mahalanobis distance calculation so that the number of pixels sampled is not limited by the number of training images.

We evaluated our system with over 200 clinical images from prostate and breast data, of multiple modalities. Expert radiologists were recruited to perform manual segmentations, and a 10-fold cross validation was performed to evaluate, via 6 distinct metrics, the effectiveness and accuracy of our system. The rest of the paper will be organized as follows. The methodology will be presented in Section 2, which will be followed by a thorough description of all data sets and all

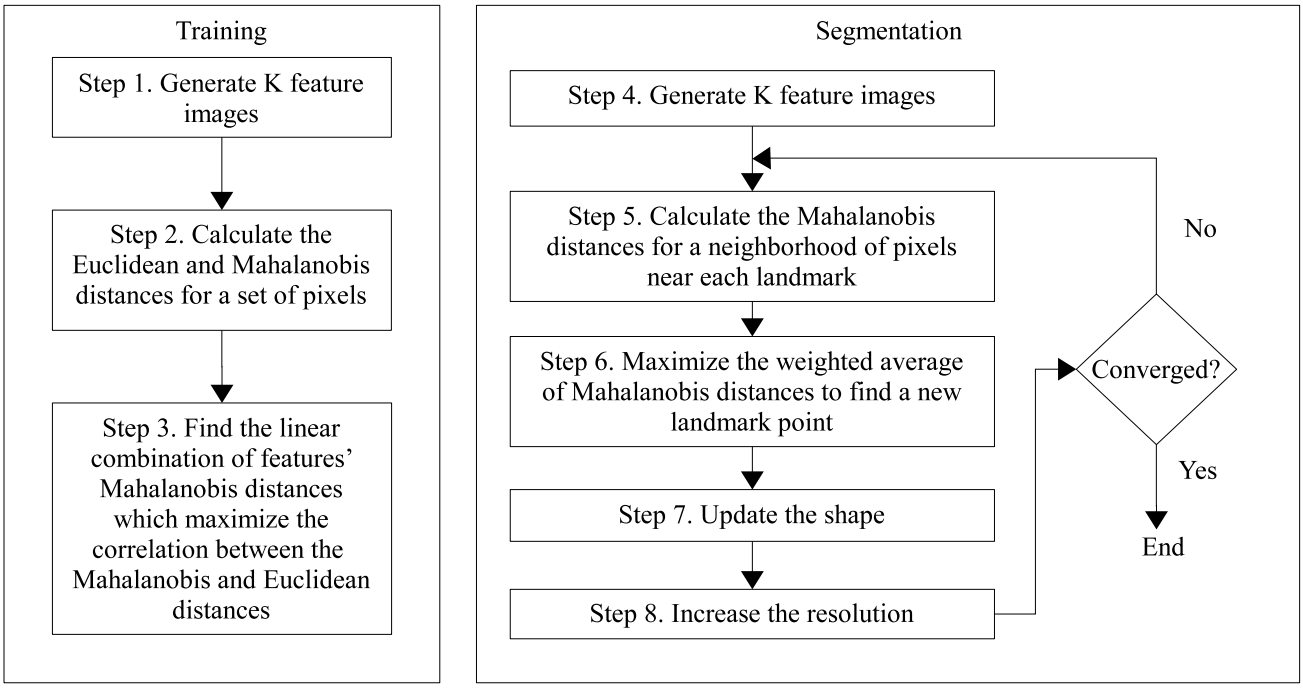

Figure 1. The modules and pathways comprising WERITAS with the training module on the left and the segmentation module on the right.

experiments performed on those data sets in Section 3. Then the results and a discussion will be presented in Section 4, and finally the concluding remarks will be presented in Section 5. The methodology section will first define the notation, and will be followed by the generation of features, the intelligently selection of features, and will conclude with the actual segmentation of a new image.

2. METHODOLOGY

Notation

We define the set of N training images as $S_{tr} = \{\mathcal{C}^n \mid n \in \{1, \ldots, N\}\}$, where $\mathcal{C}^n = (C, g^n)$ is an image scene where $C \in \Re^2$ represents a set of 2D spatial locations and $g^n(c)$ represents a function that returns the intensity value at any $c \in C$. We define a feature scene as $\mathcal{F}_k = (C, g_k)$ where $g_k(c)$ contains the feature value at pixel $c \in C$. Finally, we define a κ-neighborhood centered on c as $\mathcal{N}_\kappa(c)$, where for $\forall d \in \mathcal{N}_\kappa(c), \parallel d - c \parallel_2 \leq \kappa$.

2.1 Brief Overview of WERITAS

The steps in this section correspond to the steps in Figure 1.

Training

Step 1. A large set of K texture features are generated for all training images. For our implementation, these include gradient filters, and 1st and 2nd order statistical texture features.

Step 2. A set of pixels are sampled, and the Mahalanobis distance to the training data and the Euclidean distance to the ground truth are calculated for all pixels. This allows the correlation between the 2 distances to be determined for each feature.

Step 3. The features' Mahalanobis distances are linearly combined to maximize the correlation coefficient. This results in a set of features (and corresponding weights) which yield the highest correlation between Euclidean and Mahalanobis distance when linearly combined.

Segmentation

Step 4. The same K texture features from Step 1 are generated for a test image.

Step 5. A region is search near each landmark for the object border, and each potential pixel has a neighborhood sampled from the features selected in Step 3.

Step 6. The pixel with the lowest weighted average of Mahalanobis distances is presumed to be a better location for the object border.

Step 7. The shape model is updated to fit the newly found landmark points.

Step 8. The resolution is increased (if possible), and the process repeats at Step 5 until convergence.

2.2 Feature Extraction

For $\forall \mathcal{C}^n \in S_{tr}, K$ features scenes $\mathcal{F}_k^n = (C, g_k^n), k \in \{1, \ldots, K\}$ are then generated. These are generated by taking an image $\mathcal{C}^n$, and $\forall c \in C$, a function is performed on the vector $S_f = \{g^n(d) \mid d \in \mathcal{N}_\kappa(c)\}$. For our implementation, we used gradient,[15] first order statistical, and second order statistical features, resulting in 84 total features ($K = 84$) defined below.

Gradient Features

Sobel operators were convolved with S_f to detect the strength of the horizontal, vertical, and diagonal edges. In addition, four Kirsch linear operators were convolved with S_f to detect the strength of edges normal to lines oriented $0, \pi/4, \pi/2$, and $3\pi/4$. Finally, a set of linear Gabor kernels $Gab(x, y) = exp((x' + y')/\sigma) \cdot cos(2\pi x'/\lambda)$ where $\sigma = \{1, 3, 5\}$ and $\lambda = \{0, \pi/4, \pi/2, 3\pi/4\}$ were convolved with S_f to obtain the Gabor gradient features.

First Order Statistical Features

We obtained a series of first order statistical texture features by calculating $mean(S_f), median(S_f), mode(S_f), range(S_f)$, and $standarddeviation(S_f)$.

Second Order Statistical Features

The second order statistical texture features extracted were the Haralick features.[16] The Haralick features were calculated from a gray level square co-occurrence matrix P, where each element (u, v) of P indicates the frequency with which two distinct pixels $d, e \in \mathcal{N}_\kappa(c)$ with associated intensities $g^n(d) = u, g^n(e) = v$ are adjacent. From P, the following 16 Haralick features were calculated: Energy, Entropy, Inertia, Correlation, Inverse Difference Moment, Information Correlation 1 and 2, Sum Average, Sum Variance, Sum Entropy, Difference Average, Difference Variance, Difference Entropy, Shade, Prominence, and Variance.

2.3 Calculating the Mahalanobis Distances

We now wish to find the optimal linear combination of M features, and so a set of M features, $f_m, m \in \{1, \ldots, M\}\}, f_m \in \{1, \ldots, K\}$, and weights, $\alpha_m, m \in \{1, \ldots, M\}$, must now be selected for each landmark point, where $M < K$. For $\forall \mathcal{C}^n \in S_{tr}, c^n$ is a landmark point on the border manually delineated by an expert. For each landmark point c^n and each feature k, the feature values for $\forall d \in \mathcal{N}_\kappa(c^n)$ are denoted as the vector $\mathbf{g}_k^n = \{g_k^n(d) \mid d \in \mathcal{N}_\kappa(c^n)\}$. The mean vector over all n's is given as $\bar{\mathbf{g}}_k$ and the covariance matrix is given as Φ_k. We now take are large set of pixels $S_{px} = \{e \mid e \in C^n, n \in \{1, \ldots, N\}\}$, and sample a neighborhood around each pixel on each feature k, which is denoted as $\mathbf{g}_{k,e} = \{g_k^n(d) \mid d \in \mathcal{N}_\kappa(e)\}$. The number of pixels chosen to sample ($|S_{px}|$) must be large enough to encompass most of the image, or at least a large region near the current landmark point, so that the chosen features will actually perform well in a real segmentation task. The set of Euclidean distances to the ground truth are then calculated as

$$E = \{\| e - c^n \|_2 \quad | e \in S_{px}\}. \tag{1}$$

The set Mahalanobis distances to the training data are then calculated for each feature as

$$\psi_k = \{(\mathbf{g}_{k,e} - \bar{\mathbf{g}}_k)^T \cdot (\Phi_k)^{-1} \cdot (\mathbf{g}_{k,e} - \bar{\mathbf{g}}_k) \mid e \in S_{px}\}. \tag{2}$$

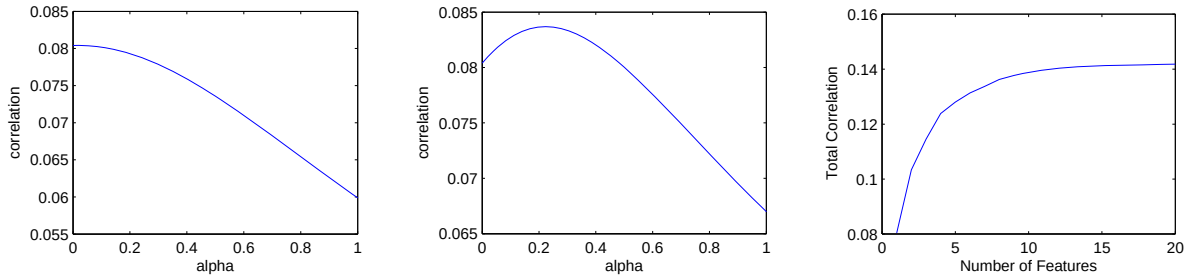

Figure 2. For (a) and (b), the X axis represents the α values, and the corresponding correlation values for a given feature are shown on the Y axis. (a) shows a feature in which $\hat{\alpha} = 0$; i.e. the correlation is not improved at all by including that feature. (b) shows a feature in which $\hat{\alpha} = 0.3$, with the correlation for that feature being 0.084. (c) shows the total correlation between Euclidean distances and the linear combination of Mahalanobis distances as each f_m and α_m are calculated.

A problem arises in Equation 2 when $(\Phi_k)^{-1}$ is undefined. Specifically, when limited training data exists (where $N < |\mathcal{N}_\kappa|$), the determinant of Φ_k would be zero, yielding an undefined Mahalanobis distance. We overcome this limitation by calculating the Moore-Penrose pseudo-inverse.[14] The Moore-Penrose pseudo-inverse can be calculated for any matrix, including matrices with a determinant of 0, which would happen if $N < |\mathcal{N}_\kappa|$, and holds most of the properties of the traditional inverse. This allows the Mahalanobis distance to be estimated without setting any requirements for the values of κ and N.

2.4 Selecting Features

There now exists a set Euclidean distances to the ground truth, E, and a corresponding set of Mahalanobis distances ψ_k for each feature $k \in \{1, \ldots, K\}$, where $|E| = |\psi_k| = |S_{px}|$. The basic metric used to determine the accuracy of a given feature will be the Pearson correlation coefficient[17] (φ), defined as

$$\varphi(\mathbf{X}, \mathbf{Y}) = \frac{1}{|\mathbf{X}| - 1} \sum_{i=1}^{|\mathbf{X}|} \left(\frac{X_i - \mu_{\mathbf{X}}}{\sigma_{\mathbf{X}}} \right) \left(\frac{Y_i - \mu_{\mathbf{Y}}}{\sigma_{\mathbf{Y}}} \right), \tag{3}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are vectors, and μ and σ represent mean and standard deviation respectively. Ideally there would be a strong correlation between the Mahalanobis distance to the training data, and the Euclidean distance to the landmark point, so that the Mahalanobis distance decreases as the distance to the ground truth decreases. This will allow the system to become closer to the ground truth by locating the minimum Mahalanobis distance. So we now wish to find the linear combination of features which maximize the correlation between the Euclidean distances (E), and the resulting linear combination of Mahalanobis distances. The best feature is given as the feature with the highest correlation; that is, $f_1 = \text{argmax}_k [\varphi(E, \psi_k)]$, and we set $\alpha_1 = 1$. Features $f_m, m \in \{2, \ldots, M\}$ are determined as follows. First, the optimal α is determined for each potential feature k, denoted as $\hat{\alpha}_k$. The $\hat{\alpha}_k$ value will be the weighting of feature k which maximizes the correlation coefficient given the already-selected features $\{f_1, \ldots, f_{m-1}\}$ and corresponding weights $\{\alpha_1, \ldots, \alpha_{m-1}\}$. This correlation is calculated from a weighted average of the Mahalanobis distances of all the previously selected features, so that

$$\hat{\alpha}_k = \underset{\alpha}{\text{argmax}} \left[\varphi \left(E, \psi_{f_1} \cdot \alpha_1 + \ldots + \psi_{f_{m-1}} \cdot \alpha_{m-1} + \psi_k \cdot \alpha \right) \right]. \tag{4}$$

Therefore, each $\hat{\alpha}_k$ denotes the α value that would increase the correlation the most for feature k. Several values of α and φ are shown for 2 different features in Figures 2(b) and 2(a). Now the feature which increased the correlation the most (given its optimal $\hat{\alpha}_k$) is chosen, so that

$$f_m = \underset{k}{\text{argmax}} \left[\varphi(E, \psi_{f_1} \cdot \alpha_1 + \ldots + \psi_{f_{m-1}} \cdot \alpha_{m-1} + \psi_k \cdot \hat{\alpha}_k) \right] \tag{5}$$

and $a_m = \hat{\alpha}_{f_m}$. This will result in a linear combination of features which maximizes the correlation between Euclidean and Mahalanobis distances. The φ values are shown in Figure 2(c) as each f_m is calculated. It can be seen that after adding about 15 features, the correlation does not significantly increase as more features are added. The algorithm below shows how our algorithm works.

Algorithm $FeatureSelect$
Input: $M, (N \cdot K)$ Feature images, N landmarks c^n.
Output: $f_m, \alpha_m, m \in \{1, \ldots, M\}, f_m \in \{1, \ldots, K\}$
$begin$
0. Calculate $\bar{\mathbf{g}}_k$ and Φ_k from the N images for $\forall k \in \{1, \ldots, K\}$;
1. Calculate ψ_k and E for a large number of pixels S_{px} (Equations 1 and 2) for $\forall k \in \{1, \ldots, K\}$;
2. for $m = 1$ to M do
3. $\quad for$ $k = 1$ to K do
4. $\quad\quad$ Calculate the optimal $\hat{\alpha}_k$ by maximizing φ (Equation 4);
5. $\quad endfor$;
6. $\quad$ Calculate f_m by maximizing φ, given all $\hat{\alpha}_k$'s (Equation 5);
7. $\quad$ Let $\alpha_m = \hat{\alpha}_{f_m}$;
8. $endfor$;
end

2.5 Using the Selected Features for an ASM Based Segmentation

The features $f_m, m \in \{1, \ldots, M\}$ are then used in a segmentation system to segment a new test image $\mathcal{C}_{te}$ where $\mathcal{C}_{te} \notin S_{tr}$. First, the K features are calculated from $\mathcal{C}_{te}$, denoted as $\mathcal{F}_k = (C, g_k)$. The traditional ASM approach is used: find new landmark points, and then update the shape to fit these points. Since this paper focuses on the landmark border detection, we will leave the shape updating algorithm the same as in traditional ASM's.[1] For iteration i, we denote the current landmark as l_i, so that the goal is to find l_{i+1}. To find l_{i+1}, a κ-neighborhood is searched near l_i, denoted as $\mathcal{N}_\kappa(l_i)$. For $\forall \hat{l} \in \mathcal{N}_\kappa(l_i)$, where $\hat{l}$ is a potential new landmark point, a κ-neighborhood is sampled, denoted as $\mathcal{N}_\kappa(\hat{l})$. Then, each feature's values are sampled from this neighborhood, denoted as $\mathbf{g}_{k,\hat{l}} = \{g_k(d) \mid d \in \mathcal{N}_\kappa(\hat{l})\}$. Finally, l_{i+1} is found by maximizing the weighted average of the Mahalanobis distance for the M features, so that

$$l_{i+1} = \underset{\hat{l} \in \mathcal{N}_\kappa(l_i)}{\operatorname{argmax}} \left[\sum_{m=1}^{M} \left(\mathbf{g}_{f_m,\hat{l}} - \bar{\mathbf{g}}_{f_m} \right)^T \cdot \left(\Phi_{f_m} \right)^{-1} \cdot \left(\mathbf{g}_{f_m,\hat{l}} - \bar{\mathbf{g}}_{f_m} \right) \cdot \alpha_m \right]. \tag{6}$$

This is the same linear combination of Mahalanobis distances which maximized the correlation between Euclidean and Mahalanobis distance. Ideally, if there was a strong correlation between the Euclidean and the linear combination of Mahalanobis distances, then l_{i+1} would be closer to the true landmark point after each iteration. Once a set of new landmarks have been selected, the shape is updated, and the system moves on to the next iteration. The algorithm below describes exactly how the segmentation works.

Algorithm $ASMFeatureSelect$
Input: A set of initial landmark points, each one denoted as l_1, K feature images,
M features and weights $f_m, \alpha_m, m \in \{1, \ldots, M\}, f_m \in \{1, \ldots K\}$
Output: Final set of landmark points
$begin$
00. Initialize $i = 0$
01. $while\ \forall l_i \neq l_{i+1}$
02. Let $i = i + 1$;
03. $for\ each\ \hat{l} \in \mathcal{N}_\kappa(l_i)$ do
04. $for\ m = 1$ to M do
05. Sample a pixels from a neighborhood surrounding $\hat{l}$ on feature image f_m, denoted as $\mathbf{g}_{f_m,\hat{l}}$;
06. $endfor$;
07. $endfor$;
08. Calculate l_{i+1} by maximizing the weighted average of Mahalanobis distances (Equation 6);
09. Repeat Steps 3 - 8 for all landmark points;
10. Fit the statistical shape model to the new landmark points;
11. Increase the image resolution;
12. $endwhile$
end

3. EXPERIMENTS AND EVALUATION METHODS

3.1 Experimental Setup

For every image, an expert radiologist was used to manually segment the tissue border, which was used as the ground truth for training and for evaluating the effectiveness of the automatic segmentation. All experiments were performed using a 10-fold cross validation as follows. A set of images from a single data set was randomly split up into 10 groups. To test each group, the other 9 groups were used to train, and the segmentation system was performed on the images in the current group. This was repeated for all 10 groups, and the resulting mean and standard deviation values were then calculated. The system was tested on the 3 different data sets summarized in Table 1.

Table 1. Description of the data sets used to evaluate the segmentation system.

Notation	Tissue	Protocol	Field Strength	# of Slices	Resolution (pixels)	Resolution (mm)
D_1	Prostate	T2-W MRI	1.5T	128	256×256	140×140
D_2	Prostate	T2-W MRI	3T	83	512×512	140×140
D_3	Breast	DCE	1.5T	10	512×512	140×140

Each image was initialized by placing the mean shape into the center third of the image, and the entire system was performed in a multi-resolution fashion, starting at 32×32 pixels. The neighborhoods $\mathcal{N}_\kappa$ were set as circles with radius of 5 pixels ($\kappa = 5$).We started with 84 features ($K = 84$), and selected 5 features ($M = 5$). For D_1 and D_2 (D_3 did not have enough slices for quantitative analysis), we tested the traditional ASM, WERITAS, and an ASM using random features, all of which are summarized as follows.

- Experiment 1: ASM. This experiment is the traditional ASM, in which $f_m = (\text{intensities}), \alpha_m = 1, m = \{1\}$.
- Experiment 2: MANTRA. The outline for averaging the Mahalanobis distance for a set of random features was presented in the MANTRA paper,[9] so the MANTRA experiment consists of the following: $f_m = random(\{1, \ldots, K\})$, $\alpha_m = 1, m \in \{1, \ldots, 5\}$.
- Experiment 3: WERITAS. Our new $FeatureSelect$ algorithm with $M = 5$ is used to choose features f_m and weights α_m for this experiment.

3.2 Quantitative Evaluation

The quantitative metrics utilized were overlap, sensitivity, specificity, positive predictive value (PPV), mean absolute distance error (MAD), and Hausdorff distance error.[10,18] We define a resulting automatic segmentation of the border of image $\mathcal{C}$ as $G_1^a(\mathcal{C})$, the set of pixels inside $G_1^a(\mathcal{C})$ as $G_2^a(\mathcal{C})$. The corresponding manual expert segmentation for image $\mathcal{C}$ is denoted as $G_1^m(\mathcal{C})$ for the border pixels, with $G_2^m(\mathcal{C})$ denoting the pixels inside the border. We now define our metrics as follows. First, for the area based metrics (overlap, sensitivity, specificity, and PPV), we define the number of true positive pixels as $TP = |G_2^a(\mathcal{C}) \cap G_2^m(\mathcal{C})|$, the number of true negative pixels as $TN = |\mathcal{C} - G_2^a(\mathcal{C}) - G_2^m(\mathcal{C})|$, the number of false positive pixels as $FP = |G_2^a(\mathcal{C}) - G_2^m(\mathcal{C})|$, and the number of false negative pixels as $FN = |G_2^m(\mathcal{C}) - G_2^a(\mathcal{C})|$. We now define our metrics as follows (where 1 - 4 are area based and 5 - 6 are edge based).

1. Overlap = $TP/(FP + TP + FN)$.
2. Sensitivity = $TP/(TP + FN)$.
3. Specificity = $TN/(TN + FP)$.
4. PPV = $TP/(TP + FP)$.
5. MAD = $mean_d\,[min_e \parallel d - e \parallel_2 \quad | \, e \in G_1^m(\mathcal{C}), d \in G_1^a(\mathcal{C})]$.
6. Hausdorff distance = $max_d\,[min_e \parallel d - e \parallel_2 \quad | \, e \in G_1^m(\mathcal{C}), d \in G_1^a(\mathcal{C})]$.

The mean and standard deviation values from each of these metrics were then run through a Student's t-test to determine the statistical significance of all results, with the resulting p values reported.

4. RESULTS AND DISCUSSION

4.1 Quantitative Results

Data Set #1 (D_1) - 1.5T *in vivo* MR Prostate Images

The mean and standard deviation results from the cross validation are shown in Table 2. For this data set, to test inter-expert variability, a second expert radiologist was asked to also segment the images, with that expert's segmentations denoted as "Expert 2". In addition, a student's t-test was performed to determine the significance of all results, and the results p-values are shown in Table 3. It can be seen that WERITAS with $M = 5$ features performed significantly better in most metrics than the traditional ASM and MANTRA with random features, which was confirmed by a t-test ($p < 0.05$ for most results). The only metric in which MANTRA outperformed WERITAS (specificity) was not statistically significant. The specificity values were extremely high in all cases due to the fact that the region of the image containing the prostate was so small compared to the size of the image, yielding an unreasonably high TN value. In addition, the fact that the random features did better than intensities indicates that for this data set, the feature images in general contained more useful information than the intensities. However, the fact that WERITAS did not always give very significant differences from MANTRA (with random features) indicates that our original set of K features did not contain any features that could perform extremely well. There is also a high correlation between the Overlap results from MANTRA and the Overlap results from WERITAS (correlation = 0.73), indicating that both methods performed poorly on the same set of images, suggesting that those images were particularly difficult to segment regardless of which features were used.

Data Set #2 (D_2) - 3T *in vivo* MR Prostate Images

D_2 shows a similar trend as D_1, in that the new WERITAS method performs better than both the traditional ASM and MANTRA with random features, which can be seen in Table 4 and shown statistically in Table 5. The traditional ASM outperformed WERITAS in the specificity and PPV metrics, but not to any statistical significance. It's interesting to note that the traditional ASM outperformed the MANTRA in most of the metrics, indicating that in this data set, the intensities were better than a plethora of random features. However, in almost all cases, an intelligent selection of features (WERITAS) was shown to perform the best, supporting the idea of maximizing the correlation between Euclidean and Mahalanobis distance to select features.

Table 2. Quantitative results for D_1 in terms of mean $\pm$ standard deviation from the 10-fold cross validation for the traditional ASM, the ASM using random features (MANTRA), WERITAS, and a second expert radiologist (Expert 2). The best results for each metric are shown in bold.

Experiment	Area Based Metrics				Edge Based Metrics (mm)	
	Overlap	Sensitivity	Specificity	PPV	MAD	Hausdorff
ASM	$.662 \pm .154$	$.824 \pm .159$	$.996 \pm .004$	$.810 \pm .192$	1.851 ± 1.235	4.701 ± 3.082
MANTRA	$.815 \pm .074$	$.896 \pm .074$	$.999 \pm .002$	$\mathbf{.909 \pm .093}$	$1.665 \pm .793$	5.189 ± 2.331
WERITAS	$\mathbf{.833 \pm .069}$	$\mathbf{.925 \pm .058}$	$\mathbf{.999 \pm .001}$	$.898 \pm .082$	$\mathbf{1.509 \pm .751}$	$\mathbf{4.642 \pm 2.093}$
Expert 2	$.858 \pm .101$	$.886 \pm .083$	$.999 \pm .001$	$.961 \pm .089$	$1.246 \pm .807$	4.455 ± 2.571

Table 3. Statistically significant results for D_1 between the traditional ASM, MANTRA, and WERITAS. One asterisk indicates $p < 0.05$, two indicates $p < 0.01$, three indicates $p < 0.001$, and four indicates $p < 0.0001$.

	ASM / MANTRA	ASM / WERITAS	MANTRA / WERITAS
Area Metrics	Overlap**** Sensitivity**** Specificity**** PPV****	Overlap**** Sensitivity**** Specificity**** PPV****	Overlap* Sensitivity***
Edge Metrics	Hausdorff**** MAD****	Hausdorff**** MAD****	Hausdorff*

Table 4. Quantitative results for D_2 in terms of mean $\pm$ standard deviation from the 10-fold cross validation for the traditional ASM, the ASM using random features (MANTRA), and WERITAS. The best results for each metric are shown in bold.

Experiment	Area Based Metrics				Edge Based Metrics (mm)	
	Overlap	Sensitivity	Specificity	PPV	MAD	Hausdorff
ASM	$.592 \pm .129$	$.693 \pm .183$	$\mathbf{.991 \pm .014}$	$\mathbf{.869 \pm .169}$	4.114 ± 1.630	9.130 ± 3.712
MANTRA	$.544 \pm .142$	$.836 \pm .081$	$.966 \pm .020$	$.623 \pm .192$	6.164 ± 1.354	14.46 ± 2.948
WERITAS	$\mathbf{.765 \pm .108}$	$\mathbf{.908 \pm .064}$	$.988 \pm .012$	$.840 \pm .140$	$\mathbf{2.451 \pm 1.587}$	$\mathbf{6.900 \pm 3.950}$

Table 5. Statistically significant results for D_2 between the traditional ASM, MANTRA, and WERITAS. One asterisk indicates $p < 0.05$, two indicates $p < 0.01$, three indicates $p < 0.001$, and four indicates $p < 0.0001$.

	ASM / MANTRA	ASM / WERITAS	MANTRA / WERITAS
Area Metrics	Overlap* Sensitivity**** Specificity**** PPV****	Overlap**** Sensitivity****	Overlap**** Sensitivity**** Specificity**** PPV****
Edge Metrics	Hausdorff**** MAD****	Hausdorff*** MAD****	Hausdorff**** MAD****

4.2 Qualitative Results

Figure 3 shows a side by side comparison of the ASM system (left column), MANTRA (middle column), and WERITAS (right column). It can be seen that in the case of D_1 (rows 3 and 4), both MANTRA and WERITAS perform well, but WERITAS performs slightly better, which is further supported by the quantitative results. In addition, in almost all the results it can be seen that the ASM system typically undersegments, while the random features typically oversegment the image. There is at least one representative image from the base, midgland, and apex of the prostate in rows 4, 3, and 2 respectively, showing that the advantages of WERITAS are not restricted to one region of the gland. Finally, the results look more jagged in the images from D_1 due to the fact that the image resolution is lower than that of D_2, yielding a coarser segmentation. Finally, it can be seen in Row 5 that a simple task, such as that presented in data set D_3 yields excellent segmentations regardless of which features are or aren't used, suggesting that the differences between methods are more visible in difficult segmentation tasks.

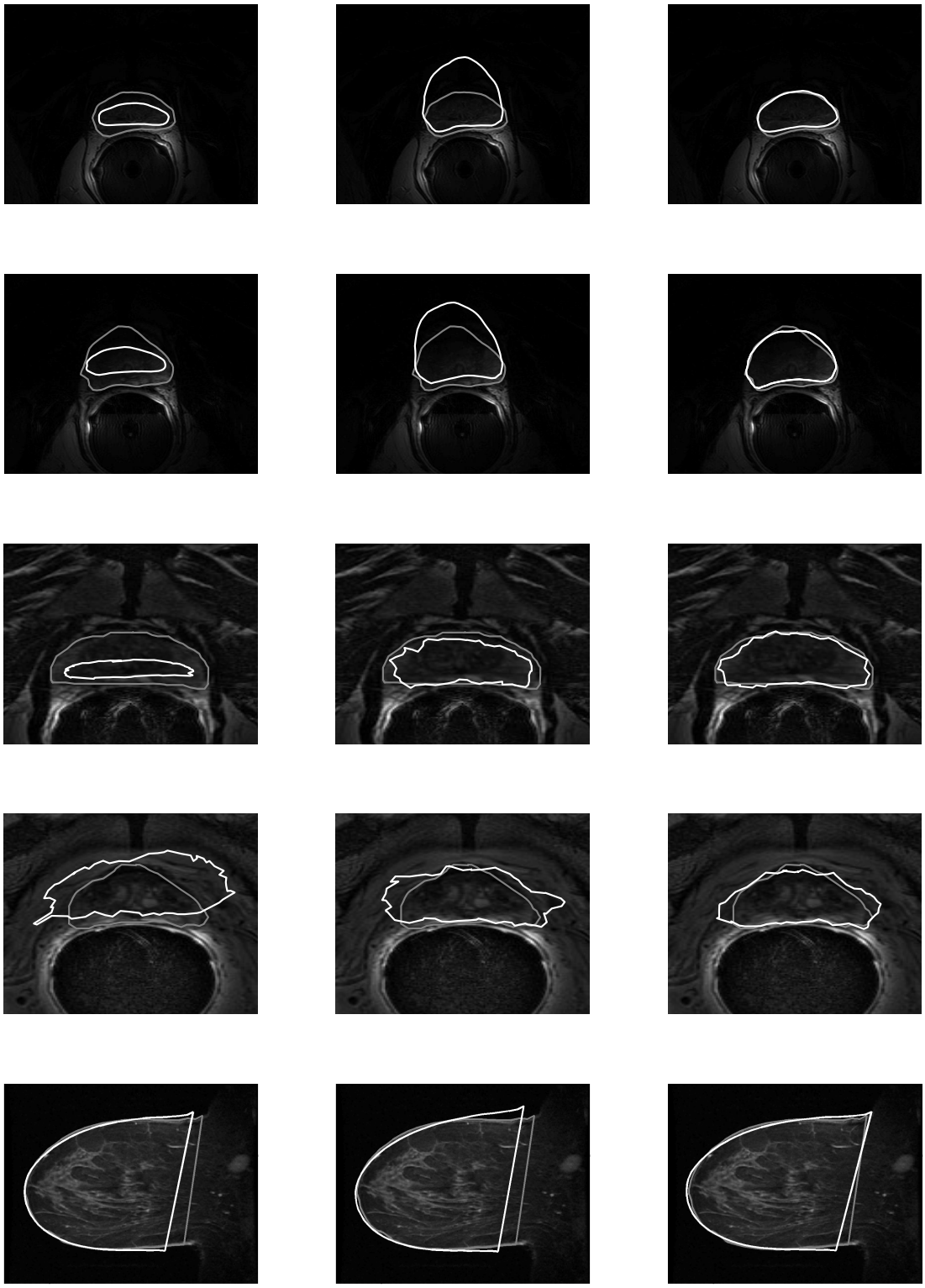

Figure 3. Shown here are qualitative results with the ground truth in gray, and the resulting segmentation in white. The left column shows the result from the ASM system, the middle column shows the result from MANTRA (random features), and the right column shows the result from WERITAS. Rows 1 and 2 show images from D_2, and rows 3 and 4 show images from D_1, and Row 5 from D_3.

5. CONCLUDING REMARKS

We have presented WERITAS, a novel segmentation system which is based on the popular ASM segmentation system. WERITAS takes a weighted average of Mahalanobis distances to more accurately detect the object border. The features are selected in an intelligent fashion, maximizing the correlation between the Euclidean distance and the Mahalanobis distance. The motivation behind maximizing the correlation is that when there is a high correlation, the locations with a low Mahalanobis distance will also be closer to the ground truth, yielding a more accurate segmentation system. We have evaluated our approach on over 200 clinical images via 6 metrics, and shown that our system is statistically better than using the traditional approach with only intensities and statistically better than using a random set of features.

ACKNOWLEDGMENTS

Work made possible via grants from Coulter Foundation (WHCF 4-29368), New Jersey Commission on Cancer Research, National Cancer Institute (R21CA127186-01, R03CA128081-01), and the Society for Imaging Informatics in Medicine (SIIM). The authors would like to acknowledge the ACRIN database for the MRI data.

REFERENCES

[1] Cootes, T., Taylor, C., Cooper, D., and Graham, J., "Active shape models - their training and application," *Computer Vision and Image Understanding* **61**, 38–59 (Jan 1995).

[2] Cootes, T., Edwards, G., and Taylor, C., "Active appearance models," in [*ECCV '98*], 484–498 (1998).

[3] Beichel, R., Bischof, H., Leberl, F., and Sonka, M., "Robust active appearance models and their application to medical image analysis," *IEEE Trans Med Imag* **24**, 1151–1170 (Sep. 2005).

[4] Robers, M. and Graham, J., "Robust active shape model search," in [*European Conference on Computer Vision*], 517–530 (2002).

[5] Cootes, T., Taylor, C., and Lanitis, A., "Multi-resolution search with active shape models," *Computer Vision and Image Processing* **1**, 610–612 (Oct 1994).

[6] Toth, R., Tiwari, P., Rosen, M., Madabhushi, A., Kalyanpur, A., and Pungavkar, S., "An integrated multi-modal prostate segmentation scheme by combining magnetic resonance spectroscopy and active shape models," in [*SPIE Medical Imaging*], **6914**(1) (2008).

[7] van Ginneken, B., Frangi, A., Staal, J., Romeny, B., and Viergever, M., "Active shape model segmentation with optimal features.," *IEEE Trans Med Imag* **21**, 924–933 (Aug 2002).

[8] Seghers, D., Loeckx, D., Maes, F., Vandermeulen, D., and Suetens, P., "Minimal shape and intensity cost path segmentation.," *IEEE Trans Med Imag* **26**, 1115–1129 (Aug 2007).

[9] Toth, R., Chappelow, J., Rosen, M., Pungavkar, S., Kalyanpur, A., and Madabhushi, A., "Multi-attribute, non-initializing, texture reconstruction based asm (mantra)," in [*MICCAI*], *Lecture Notes in Computer Science* **1**, 653–661 (2008).

[10] Madabhushi, A., Feldman, M., Metaxas, D., Tomaszeweski, J., and Chute, D., "Automated detection of prostatic adenocarcinoma from high-resolution ex vivo mri," *IEEE Trans Med Imag* **24**, 1611–1625 (Dec 2005).

[11] Doyle, S., Madabhushi, A., Feldman, M., and Tomaszewski, J., "A boosting cascade for automated detection of prostate cancer from digitized histology," in [*MICCAI*], *Lecture Notes in Computer Science* **4191**, 504–511 (2006).

[12] Viswanath, S., Rosen, M., and Madabhushi, A., "A consensus embedding approach for segmentation of high resolution in vivo prostate magnetic resonance imagery," in [*SPIE*], (2008).

[13] Chappelow, J., Madabhushi, A., Rosen, M., Tomaszeweski, J., and Feldman, M., "A combined feature ensemble based mutual information scheme for robust inter-modal, inter-protocol image registration," in [*Biomedical Imaging: From Nano to Macro, 2007. ISBI 2007. 4th IEEE International Symposium on*], 644–647 (Apr 2007).

[14] Penrose, R., "A generalized inverse for matrices," in [*Proceedings of the Cambridge Philosophical Society*], (51) (1955).

[15] Ballard, D. and Brown, C., "Computer vision," tech. rep., Prentice Hall Professional Technical Reference (1982).

[16] Haralick, R., Shanmugam, K., and Dinstein, I., "Textural features for image classification," *IEE Transactions on Systems, Man and Cybernetics* **3**(6), 610–621 (1973).

[17] Edwards, A., [*An Introduction to Linear Regression and Correlation*], ch. 4, 33–46 (1976).

[18] Madabhushi, A. and Metaxas, D. N., "Combining low-, high-level and empirical domain knowledge for automated segmentation of ultrasonic breast lesions," *IEEE Trans Med Imag* **22**, 155–170 (Feb. 2005).