Optimal Feature Combination for Automated Segmentation of Prostatic Adenocarcinoma from High Resolution MRI

Anant Madabhushi*, Michael Feldman†, Dimitris Metaxas‡, Deborah Chute‡ and John Tomaszeweski†
*Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, USA
†Department of Surgical Pathology, Hospital at the University of Pennsylvania, Philadelphia, PA, USA
‡Departments of Computer Science and Bioengineering, Rutgers the State University of New Jersey, Piscataway, NJ, USA.

Abstract—In spite of the high global incidence of prostate cancer, limited computer-aided tools to assist in its detection exist and that too only for ultrasound images. In this work we present a novel feature ensemble scheme for combining different 3D texture features for automated segmentation of prostatic adenocarcinoma from 4T MR images. The first step of our methodology comprises of a feature extraction module to extract 3D statistical, gradient and Gabor texture features at multiple scales and orientations and generate the corresponding Feature Scenes. Every voxel in each of the Feature Scenes is assigned a likelihood of malignancy using a Bayesian inference module. These results are then combined using a novel weighted linear combination scheme; weights being determined by minimization of a cost function. The method was found to be optimal compared to other popular ensemble methods such as Boosting, Majority Rule, Product Rule and Averaging in terms of Sensitivity and Positive Predictive Value (PPV). In fact, our feature ensemble scheme also outperformed an expert radiologist in terms of Sensitivity. An interesting result from the comparison of the different feature ensembles was that Boosting performs poorly on MR data that has been corrected for background inhomogeneity.

I. INTRODUCTION

Prostatic adenocarcinoma is the most common malignancy of men with an estimated 189,000 new cases in the USA in 2002. Prostate cancer is most curable when it is detected early. While prostatic tumors with volumes less than 5-6 \text{mm}^3 cannot be detected on ultrasound, in vivo studies have shown that nodules of carcinoma \geq 3 \text{mm}^3 can be clinically detected using 1.5T magnets [1]. Use of the 4T magnet has allowed for greater resolution of sub-macrosopic findings, features that are not easily identifiable at 1.5T. The motivation for our work comes from the prospect of creating accurate and successful computer-aided detection (CAD) techniques that could show the advantages of using MRI over ultrasound.

Visual identification of prostatic adenocarcinoma is confined by presence of benign features, having overlapping texture and intensity appearance. It can appear as a defined nodule or as a less subtle infiltrate without easily discernible borders. Further, prostatic tumors can vary greatly in size, from large tumors to microscopic ones. The motivation behind this work was to develop computer-aided detection (CAD) tools that could serve as an aid to doctors for screening patients and potentially reduce the number of unnecessary biopsies.

The lack of structure and shape in the prostatic adenocarcinoma implies that texture operators are required for segmentation. Past work in this area has looked at classifying manually segmented prostatic tissue from 2D ultrasound slices. These approaches used either first or second order statistical texture features [2], [3]. Texture features that could discriminate between benign and malignant prostatic neoplasia in MR however, have as yet not been identified. Further, 2D texture operators are unable to exploit the orientation or anisotropic information latent along the z-axis.

To build a system that can discriminate textures at least as well as humans do, we need to take into account both first and second-order statistics. Gradient operators have been shown to characterize micro-textures well and it has been suggested that they show more consistent behavior as a descriptor of pathologies than co-occurrence matrices [4]. While the 2D Gabor filter has been widely used for pattern recognition problems [7] the 3D version has found limited usage in segmentation. In this paper we extend the 2D Gabor transform to 3D and use it to extract features at multiple scales and orientations. The use of different classes of 3D texture operators at multiple scales and orientations enables us to capture the entire range of variation in size, appearance and orientation of the adenocarcinoma.

Multiple classifier systems based on the combination of outputs of different classifiers have been widely used for pattern recognition problems [9], [10], [11]. Surprisingly, their use in medical applications has been limited. While it has been established that ensembles of classifiers are often much more accurate than the individual classifiers [10], the choice of the optimal ensemble method for a specific problem domain is not well defined. Different feature ensembles have been shown to be optimal on different datasets. Most of these studies have however compared the performance of the different ensembles on very few or no medical datasets and none of them have been evaluated on MR data. While voting techniques like Boosting have been shown to reduce classification error, they are sensitive to noise. Other ensemble methods such as the Product and Average rule make large assumptions about the features, which very often may not hold. We present a novel weighted linear combination scheme that produces optimal performance by minimizing the detection error of the prostatic tumor. We show that not only is our scheme able to outperform other popular feature ensembles, it also outperforms an expert radiologist in terms of Sensitivity.
Our work represents the first attempt at fully automated detection of prostatic adenocarcinoma from high resolution 3D MR data. The novelty of our work lies in (i) Use of a range of 3D texture operators (ii) Use of 4T MR, having far greater resolution than ultrasound data and (iii) Use of a novel machine learning technique for feature combination, resulting in higher Sensitivity than a human observer. Sections II-IV describe our methodology. In sections V-VII we present our results, discussions and conclusions.

II. FEATURE EXTRACTION

Before feature extraction, the image volume \(I\) is corrected for background inhomogeneity and subsequently standardized to account for acquisition-to-acquisition signal variations inherent in MR images [5]. On this corrected and standardized scene denoted by \(I_{\text{cs}}\), we apply 3 different classes of 3D texture operators, i.e. statistical, gradient and Gabor.

A. Statistical Texture Features

We compute both first and second-order statistics in 3D. The first order statistical features: intensity, median, standard deviation and average deviation are computed within a co-occurrence matrix of gray levels in the image, the size of the co-occurrence matrix is \(G \times G\). The entry \((i,j)\) in the matrix is the number of occurrences of the pair of gray levels \(i\) and \(j\). We extend the 2D formulation to 3D as,

\[
P_{x,y,z}(r,s,t) = \sum f(r,s,t) \delta(r,s,t - j) = j\]

where \((r,s,t), (r', s', t') \in M \times N \times L, (r', s', t') = (r + d \cos \psi \cos \phi, s + d \sin \psi \cos \phi, t + d \sin \phi + \eta)\) and \(J\) is the cardinality of the set. \(M, N, L\) are the dimensions of \(I_{\text{cs}}\), \(d\) is the displacement, \(\phi, \psi\) are the 3D orientations and \(\eta\) accounts for inter-slice distance. In our system we used \(d=1\) and \(\psi = \phi = \frac{\pi}{2}\). Five texture features as proposed by Haralick [6] were computed from the co-occurrence matrix at every voxel in the image volume; energy, entropy, contrast, homogeneity and correlation.

B. Gradient Features

We compute both the directional gradient and gradient magnitude in 3D. The directional gradient \(I_{\text{gs}}\) is computed as,

\[
I_{\text{gs}} = \frac{- \hat{Q}}{\| \hat{Q} \|}
\]

where \( \hat{Q} = \left[ \frac{\partial I_{\text{cs}}}{\partial x}, \frac{\partial I_{\text{cs}}}{\partial y}, \frac{\partial I_{\text{cs}}}{\partial z} \right][\tilde{n}_x, \tilde{n}_y, \tilde{n}_z]^T \)

\( \hat{Q} \) is a 3D scene representing summation of the directional gradients, \( \frac{\partial I_{\text{cs}}}{\partial x}, \frac{\partial I_{\text{cs}}}{\partial y}, \frac{\partial I_{\text{cs}}}{\partial z} \) correspond to the image gradients along the \(x, y, z\) axes respectively and \(\tilde{n}_x, \tilde{n}_y, \tilde{n}_z\) are the normalized derivatives. The gradient magnitude \(I_{\text{gs}}\) is \(\| \hat{Q} \|\).

C. Steerable Filters

The 3D MR volume can be regarded as the weighted sum of 3-D Gabor functions of the form,

\[
I_{\text{gs}}(x,y,z) = \sum_{\alpha=1}^{K} \sum_{\beta=1}^{L} \sum_{\gamma=1}^{M} e^{-j \frac{2\pi}{\nu_0} \left[\frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2} + \frac{z^2}{\sigma_z^2}\right]} \cos(2\pi \nu_0 x)
\]

where \(\nu_0\) is the frequency of a sinusoidal wave along the \(x\)-axis, and \(\sigma_x, \sigma_y, \sigma_z\) are the standard deviations of the Gaussian envelope along the \(x, y, z\) axes respectively. The set of self-similar Gabor filters are obtained by appropriate rotations and scalings of \(I_{\text{gs}}(x,y,z)\) through the generating function [7]:

\[
g_{mn}(x,y,z) = \alpha^m g(x', y', z'), \alpha \geq 1
\]

where \(g_{mn}(x,y,z)\) is the rotated and scaled version of the original filter, \(\alpha\) is the scale factor, \(m = 0, 1,..., N - 1\) is the current orientation index, \(N\) is the total number of orientations, \(m = 0, 1, 2, ..., M - 1\) is the current scale index, \(M\) is the total number of scales, and \(x', y', z'\) are the rotated coordinates:

\[
x' = a^m(x \cos \theta + y \sin \theta),
\]

\[
y' = a^m(-x \sin \theta + y \cos \theta),
\]

\[
z' = a^m z
\]

where \(\theta = \frac{\pi}{N}\) is the orientation, \(a = (\frac{1}{\sqrt{2}})^{(M-1)}\) where \(U_h\), \(U_l\) correspond to the upper and lower center frequencies of interest. We used a total of 18 different filter channels, corresponding to 6 orientations and 3 scales.

III. BAYESIAN CLASSIFIER

During training, an expert pathologist manually segmented out regions of cancer from 4T MR images by visually registering the MR with the histology. The manually segmented tumor regions on the MR were used as masks for generating the probability density functions (pdfs) for each of the 3D features. To each of the training data the different 3D texture operators were applied and different Feature Scenes were generated. For each voxel within the tumor mask the operator response was noted and values incremented in the corresponding feature histogram. In all 15 slices from 3 different glands were used for training. For a voxel \(x\) in the scene we used Bayesian inference [8] to assign a likelihood of malignancy based on each feature independently. For each input scene, we generated likelihood scenes corresponding to each feature.

\[
P(x \in \omega_j | f_r) = P(x \in \omega_j)P(f_r | x \in \omega_j)/P(f_r)
\]

where the \(a\)-posteriori probability of observing the class \(\omega_j\) given the feature \(f_r\) is given by \(P(x \in \omega_j | f_r)\), \(P(x \in \omega_j)\) is the \(a\)-priori probability of observing the class \(\omega_j\), \(P(f_r | x \in \omega_j)\) is the conditional density, \(P(f_r) = \sum_{\beta=1}^{C} P(f_r | x \in \omega_j)P(x \in \omega_j)\) and \(C\) refers to the number of classes.

IV. FEATURE COMBINATION

Multiple classifier systems are practical and effective solutions for difficult pattern recognition problems since they seek to exploit the variance in behavior of the base learners. They can be either (a) non-generative methods, confining themselves to a set of given well-designed base learners e.g. linear combination, product or (b) generative methods, which generate learners acting on the base-learning algorithm or on the structure of the dataset, e.g. Boosting.

Choices of decision combination methods are dictated by several factors: type of output the classifiers can produce, number of classes and availability of sufficient training data. While several experimental studies have been performed to
compare ensemble methods, general performance claims about a particular ensemble are difficult to make. These studies are often very specific to particular applications. In studies comparing voting ensembles [9, 10] on large real and artificial datasets, Boosting was found to perform the best. Mac Namee et al. [11] however found that Averaging produced better results than Boosting on regression datasets. None of these comparative studies however have looked at MR studies. Our novel feature ensemble method has the flavour of both generative and non-generative methods, in that while the scheme starts out with a set of well defined base learners, the contributions of the individual features are learnt during the training phase (similar to Boosting). Further since these contributions are learnt by minimizing the error in detecting the cancer, our method outperforms other popular ensemble schemes as well as an expert radiologist in terms of Sensitivity.

A. Our Methodology

Our ensemble scheme belongs to the class of General Ensemble Methods (GEM). Weights are assigned to each of the base learners depending on its importance in minimization of a cost function. For our problem, we seek to maximize the detection of the True Positive (TP) area of the cancer while at the same time minimizing the False Positive (FP) benign area. Hence, we define a cost $E(k)$ which is a function (Φ) of both TP and FP and seek to find weights $\lambda_\gamma^{(k)}$ that minimizes $E(k)$ for a given training sample $k$.

$$\sum_{\gamma=1}^{K} \lambda_\gamma^{(k)} f_\gamma^{(k)} = S^{(k)}; \quad E^{(k)} = \Phi(||S^{(k)} - R^{(k)}||)$$  \hspace{1cm} (7)

where $R^{(k)}$ refers to the ground truth, $S^{(k)}$ represents the combined likelihood scene generated from $K$ different classifiers $f_\gamma^{(k)}$, $S^{(k)}$ is a binary mask obtained by thresholding $S^{(k)}$. We use a brute force scheme to find the smallest $E^{(k)}$ for some combination of $\lambda_\gamma^{(k)}$ for each sample $k$. The final weights $\lambda_\gamma$ are then obtained as the average of $\lambda_\gamma^{(k)}$ over all $k$ and then used for all subsequent testing.

B. Other Feature Ensemble Schemes

Bayesian Averaging or the Basic Ensemble Method (BEM) is a special case of GEM, with all $\lambda_\gamma$=1. It utilizes the average of the independent estimates as the new estimate. BEM has been widely studied and it has been shown that it can significantly reduce the empirical risk.

Methods for voting classification algorithms, such as Majority and Boosting have been shown to very successful in improving the accuracy of certain classifiers for real and artificial datasets [9; 10]. Adaptive boosting (AdaBoost) proposed by Freund and Schapire has also been shown to significantly reduce the error of any learning algorithm that consistently generates classifiers whose performance is a little better than random guessing. Boosting forces weak classifiers to hierarchically learn harder parts of a classification problem. The scheme begins with a set of training examples all of which are equally weighted. AdaBoost generates classifiers sequentially for a certain number of trials ($T$) and at each iteration ($t$) the weights of the training dataset are changed based on the classifiers that were previously built. The final strong classifier is given as:

$$h(x) = \begin{cases} 1 & : \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2} \sum_{t=1}^{T} \alpha_t \\ 0 & : \text{otherwise} \end{cases}$$

where $h(x)$ is the final strong classifier for voxel $x$ and $\alpha_t$ are the weights at each iteration ($t$).

Majority vote is a popular method for combining classifier outputs. Let $F$=${f_1, f_2, \ldots, f_K}$ be a team of classifiers and $\Omega = \{\omega_1, \omega_2, \ldots, \omega_s\}$ a set of class labels. For a given image voxel $x$, the majority votes assigns $x$ the class label supported by the majority of the classifiers $f_\gamma$.

The Product rule assumes independence of the base features. The probability of the joint decision rule $P(x \in \omega_\beta | f)$ is equal to the product of the probabilities of each individual classifiers corresponding decision $p(f_\gamma | x \in \omega_\beta)$ and given as,

$$P(x \in \omega_\beta | f) = \prod_{\gamma=1}^{K} p(f_\gamma | x \in \omega_\beta) \quad (8)$$

V. RESULTS

A total of 30 slices from 5 glands were used for testing. The feature combination results were evaluated with the histological based ground truth as determined by an expert pathologist. We also compared our results with the manual segmentations of a human observer (expert radiologist). Fig. 1 shows the

![Fig. 1](image-url)
**QUANTITATIVE COMPARISON OF DIFFERENT ENSEMBLES FOR δ=0.5**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Large</td>
<td>Sensitivity</td>
<td>41.34</td>
<td>58.96</td>
<td>2.51</td>
<td>11.11</td>
<td>11.31</td>
<td>37.75</td>
</tr>
<tr>
<td></td>
<td>PPV</td>
<td>42.79</td>
<td>36.67</td>
<td>52.01</td>
<td>44.62</td>
<td>49.92</td>
<td>72.45</td>
</tr>
<tr>
<td>Small</td>
<td>Sensitivity</td>
<td>41.04</td>
<td>53.81</td>
<td>1.24</td>
<td>13.15</td>
<td>20.12</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>PPV</td>
<td>10.22</td>
<td>7.73</td>
<td>13.43</td>
<td>6.39</td>
<td>10.82</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 1 shows the results of an expert radiologist and 5 ensemble methods on large and small tumors, i.e. tumors that could and could not be visually detected by the radiologist, for δ=0.5. The average size of the small tumors expressed as a percentage of the slice area was only 0.24% compared to 7.29% for the large tumors. For small and large tumors, BEM and GEM performed the best in terms of Sensitivity, outperforming the expert. The expert however had a higher PPV. Among the feature ensembles, the Product rule produced the highest PPV for small and large tumors. The BEM had the lowest PPV for large tumors, while AdaBoost had the lowest PPV for small tumors. Interestingly, while the ensemble methods had significantly smaller PPV values for small tumors, there was no statistically significant difference in Sensitivity between the small and large tumors. Fig. 2 shows the ROC curves of Sensitivity vs. 1-Specificity for the different ensemble methods over the entire test set. As borne out by the area under the ROC curves in Fig. 2, GEM produced the best results, marginally outperforming BEM, the Majority voting rule was third with AdaBoost and the Product rule performing the worst.

**VI. DISCUSSION**

While the poor performance of AdaBoost was surprising, it has been shown [11] that the technique suffers from overfitting and insufficient and biased training data. Further AdaBoost tends to perform poorly on the so called minority class problem, an issue in domains where one class accounts for a significant majority of observations (typical of most medical datasets). In our datasets, the positive examples, i.e. cancer, accounted for less than 5% of the image area. A more severe problem however is that since AdaBoost tends to focus on the problematic examples, there is the risk that noisy data will get boosted. It has been shown that inhomogeneity correction tends to increase noise variance in MR images [12]. This would account for the poor performance of AdaBoost and why GEM and BEM performed the best, since they have a smoothing effect which reduces noise. Majority voting while suffering from the same noise problem did better than AdaBoost, since unlike AdaBoost the noisy examples were not given larger weights. The Product rule uses an unrealistic assumption i.e. independence of the features. This is certainly not true for the base learners we have used and hence accounts for its poor performance.

**VII. CONCLUSION**

We presented a methodology to perform fully automated segmentation of prostatic adenocarcinoma. Unlike previous semi-automated detection approaches that used statistical texture operators on 2D ultrasound slices, we used a novel ensemble of 3D statistical, gradient and Gabor features on high resolution volumetric MR data. Our weighted linear combination method (GEM) was found to be optimal compared to other feature ensembles in minimizing the detection error and it also did better than an expert radiologist in terms of Sensitivity. Another interesting result from the comparison of the feature ensembles was that AdaBoost tended to perform badly on MR data that has been corrected for background inhomogeneity. While these conclusions are no doubt linked to the choice of base learners and diversity of the training set, the initial results suggest an encouraging step in building better and more accurate CAD systems.

**REFERENCES**


