Prostate Detection Using Principal Component Analysis

Aamir Virani (avirani@stanford.edu)
CS 229 – Machine Learning – Stanford University – 16 December 2005

Introduction

During the past two decades, computed tomography (CT) has become a large component of diagnosis and treatment facilities around the country. Initially used to scan the brain, CT is now used to learn more about patients' internal systems when dealing with many pathologies, including cancer.

A CT scan produces a three-dimensional view of the inside of a body; a single scan produces a set of two-dimensional slices over the region of interest. Thus, a single measurement contains many data points that can be put together to generate a 3D model of the body, visible organs, tissues, and other features [1].

Doctors who work with cancer now use CT scans to plan radiation therapies so that treatment is localized to diseased tissue and minimally hurts the surrounding, healthy organs. However, hospital staff currently goes through each slice of a CT by hand to annotate the scan with location of organs like the heart, liver, kidneys, prostate, and more. This process takes a long time and is susceptible to variation between doctors [2].

Recent work in computer vision and medical instrumentation has attempted to deal with this problem. Image segmentation of the 2D slices groups similar tissues together using region growing or seeding [3]. Others have incorporated machine-learning techniques to pick out interesting regions. Zwiggelaar and Zhu have worked with active shape modeling [4], while August and Kanade have attempted some weakly-supervised training to accomplish similar goals [5].

In this paper, I will attempt to detect the prostate in the two-dimensional slices of a CT scan by using principal component analysis and a supervised-learning classification scheme. By training on a set of slices and human-drawn contours identifying the prostate, I hope to produce a classifier that can then take a new slice from either the same patient or another and highlight the prostate on it.

Approach

A single CT scan contains over 100 slices, and each consists of 512x512 integers, where each is a value in Hounsfield Units, a normalized measure of x-ray attenuation [2]. To work with system memory limitations, I decided to focus on only those slices containing a prostate contour. This dropped the number of usable slices per scan to around forty.

Each slice must be preprocessed. Instead of working with a huge 512^2-sized vector, I broke the slice into blocks of size 16x16. The blocks do not overlap, though this is a changeable parameter for future research. Other block sizes

were later attempted, like 8x8. These blocks were arranged into a vector format, and so the setup for machine-learning algorithms was complete. A set of 16^2-length vectors capture the features of each block in a slice. Training slices were then concatenated to generate the full training set.

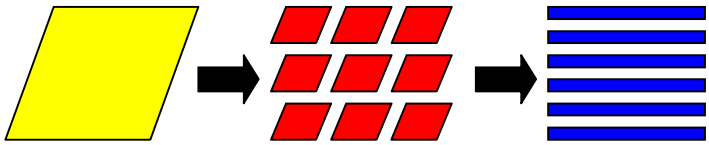

Figure 1: 2D Slice Preprocessing - go from slice to blocks to vectors

Concurrently, the related contour drawn by a doctor or other staff is processed in the same way. The resulting vectors are then compared to some threshold value. If the vector contains a certain percentage of the contour, the related slice vector is marked '1' for containing prostate. Otherwise, it is marked with a '0'.

To find a generalizable algorithm for prostate detection, principal component analysis (PCA) makes sense. Using this method should remove overfitting to our training set and the subtle effects of noise in our data. I hoped to find that organ samples clustered when projected into these lower dimensions as well. If this occurs, perhaps more can be learned about the features that differentiate organs.

In principal component analysis, one first creates the training data's covariance matrix. The eigenvectors that maximize this matrix are then used as a lower-dimensional orthogonal basis [6]. Thus, we discard the smaller variance that may be attributable to noise, differences between equipment, and changes in anatomy.

The slices in a CT scan are not independent, however. The adjacent regions in two different slices likely carry similar information and map to the same organ. To account for this, I also grouped slices and broke this into voxels that were then used in the same way as above. This three-dimensional PCA classifier was compared to the two-dimensional one.

Observations

Training a PCA-based classifier on the training set produces a set of components which most capture the variance in the signal. In the two-dimensional 16x16 block setup, the first principal component captured nearly ninety percent of the variance, while the first fifty included over ninety-nine percent.

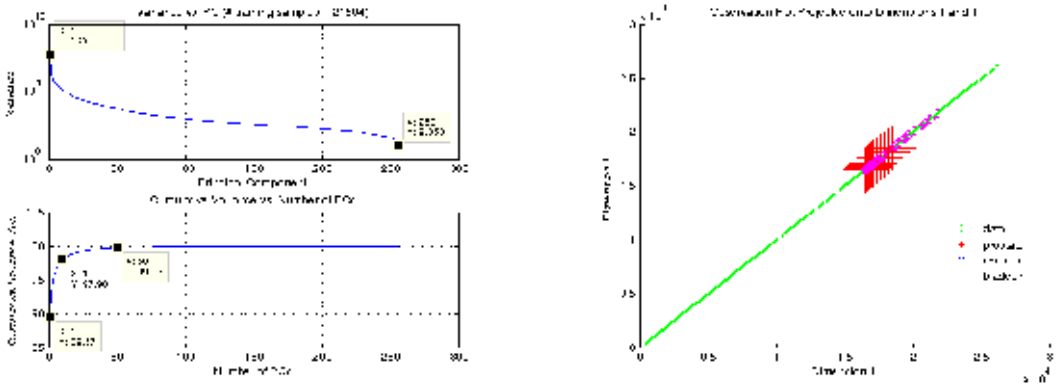

Figure 2: 2D PCA (16x16) - variance by principal components and prostate locations

The right figure above shows how the training set condenses into a one-dimensional space defined by the first principal component. Here, one can see that the prostate lies in one general region. While this may be a good sign, additional data points are plotted showing the surrounding organs often confused with the prostate, like the bladder and seminal vesicles. Unfortunately, those organs are also in the same subset of the space.

I had hoped to see soft-tissue organs separate in these lower dimensions, but this did not occur. The two-dimensional mappings below show that the PCA projections did not capture such differences. While the prostate samples did group in one region, so did the bladder and seminal vesicles. There seems to be no way to separate these three organs because they overlap in the defined space.

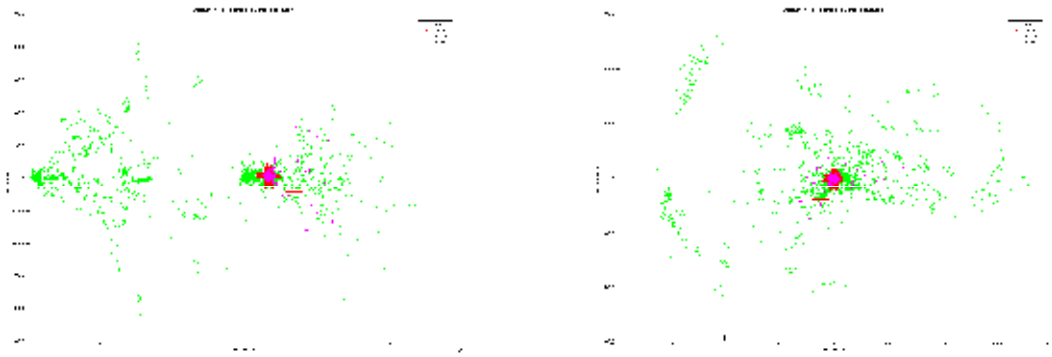

Figure 3: Clustering Observations - dimensions a) 1 and 7 and b) 2 and 4

Results

Training on half of one person's dataset yielded 20 training examples. The remaining samples of the single person were then classified, and the classifier was also run on two other patient datasets.

PCA Type	Prostate Correct (%)	Prostate Correct (%)
2D (16x16)	15	11
3D (16x16)	25	

Table 1: Correct Prostate Detection Results Using PCA

Unfortunately, the system does not work as well as hoped. However, the above results are for correct prostate detection. That is, the success rate is based on the number of prostate detections per total number of prostate blocks. This does not take into account the successful classification of non-prostate. It also does not include nearby detected blocks, such as those adjacent to the desired contour or those in between contour lobes. In other words, a few of the classifier errors may have been correct given a different doctor's contour for comparison.

The next table shows that the results are quite good if all of the data is included, but note that this is not surprising. Organs, bone, and other portions of the body have a very different makeup, so their images look quite different. Furthermore, the prostate is a very small part of the body, so classifying all blocks as non-prostate would return a very high success rate as well.

PCA Type	Overall Correct (%)
2D (16x16)	98.9
3D (16x16)	99.5

Table 2: "Overall" Classification Results Using PCA Classifier

The difficulty of this problem is when dealing with soft tissue and the lower abdomen. By looking at individual results, though, one can see that the machine-learning algorithm has correctly deduced that soft tissue is the most likely type of block containing a prostate. The classifier rarely moves out of the central portion of a slice. No edges of the body were classified as prostate, and very rarely was the inside of another organ, like a kidney, falsely classified.

Figure 4: Sample Results From PCA Classifier

In the above figure, the left image shows a typical prostate contour near the middle of the organ. It is round and centered in the body. The resulting classification performs decently. While it does not map the contour region, it does detect the surrounding area of soft tissue.

On the other hand, the right image shows the troubles of detecting a prostate shape that is not part of the original training set. When processed, one of the end lobes is detected, but the thin middle portion is missed. Note that this instance also produces a scattering of detections as well.

Future Work

Such results indicate that this is a promising direction in detection. With a larger training set that encompasses all slices of a prostate and tens or hundreds of patients, the principal vectors may generalize and produce better detection on test patients. If there were a way to selectively choose slices within in a scan, we might find that a subset of representative shapes and slices work as a better, faster training set than just using a whole chunk.

Because each CT slice was broken into equal-sized blocks, using computer vision techniques to segment the human into interesting regions before attempting to find identifiable features or components may work better. Again, current prostate segmentation techniques rely on segmentation, but incorporating machine learning to better divide soft tissue remains a nascent area of research.

Acknowledgements

I would like to thank Brian Thorndyke and the Stanford Medical School Radiation Oncology Department for their help in understanding and gathering data for this project.

References

[1] “Computed Tomography.” *Wikipedia, The Free Encyclopedia*. 18 Nov 2005, 16:42 CST. 21 Nov 2005 <http://en.wikipedia.org/wiki/Computed_tomography>.

[2] Thorndyke, Brian. Interview. 18 Oct 2005.

[3] Mazonakis, M., et. al. “Image segmentation in treatment planning for prostate cancer using the region growing technique.” British Journal of Radiology 74 (2001): 243-249.

[4] Zhu, Y., Williams, S., Zwiggelaar, R. “Segmentation of volumetric prostate MRI data using hybrid 2D+3D shape modeling.” Proceedings of Medical Image Understanding and Analysis (2004): 61-64.

[5] August, Jonas and Kanade, Takeo. “Weakly-Supervised Segmentation of Non-Gaussian Images via Histogram Adaptation.” Proceedings, 2002 Medical Image Computing and Computer Assisted Intervention (MICCAI) (2003): 992-993.

[6] Ng, Andrew. *CS229 Lecture Notes: Part 9 – Principal Components Analysis*. 16 Nov 2005.